(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 October 2001 (25.10.2001)

PCT

(10) International Publication Number
WO 01/79230 A2

(51) International Patent Classification[7]: C07H

(21) International Application Number: PCT/US01/12273

(22) International Filing Date: 13 April 2001 (13.04.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/197,514 18 April 2000 (18.04.2000) US

(71) Applicant *(for all designated States except US)*: **GENAISSANCE PHARMACEUTICALS, INC.** [US/US]; Five Science Park, New Haven, CT 06511 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **CHEW, Anne** [US/US]; 1477 Beacon Street #64, Brookline, MA 02446 (US). **CHOI, Julie, Y.** [US/US]; 38 Elizabeth Street, West Haven, CT 06516 (US). **KOSHY, Beena** [IN/US]; Apartment 11B, 1298 Hartford Turnpike, North Haven, CT 06473 (US). **ROUNDS, Eileen** [US/US]; 40 Gold Star Road, Cambridge, MA 02140 (US).

(74) Agents: **FIELD, Gisela, M.** et al.; Genaissance Pharmaceuticals, Inc, Five Science Park, New Haven, CT 06511 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/79230 A2

(54) Title: HAPLOTYPES OF THE UGT1A1 GENE

(57) Abstract: Novel single nucleotide polymorphisms in the human UDP glycosyltransferase 1 (UGT1A1) gene are described. In addition, various genotypes, haplotypes and haplotype pairs for the UGT1A1 gene that exist in the population are described. Compositions and methods for haplotyping and/or genotyping the UGT1A1 gene in an individual are also disclosed. Polynucleotides containing one or more of the UGT1A1 polymorphisms disclosed herein are also described.

HAPLOTYPES OF THE UGT1A1 GENE

RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application Serial No. 60/197,514 filed April 18, 2000.

FIELD OF THE INVENTION

This invention relates to variation in genes that encode pharmaceutically-important proteins. In particular, this invention provides genetic variants of the human UDP glycosyltransferase 1 (UGT1A1) gene and methods for identifying which variant(s) of this gene is/are possessed by an individual.

BACKGROUND OF THE INVENTION

Current methods for identifying pharmaceuticals to treat disease often start by identifying, cloning, and expressing an important target protein related to the disease. A determination of whether an agonist or antagonist is needed to produce an effect that may benefit a patient with the disease is then made. Then, vast numbers of compounds are screened against the target protein to find new potential drugs. The desired outcome of this process is a lead compound that is specific for the target, thereby reducing the incidence of the undesired side effects usually caused by activity at non-intended targets. The lead compound identified in this screening process then undergoes further *in vitro* and *in vivo* testing to determine its absorption, disposition, metabolism and toxicological profiles. Typically, this testing involves use of cell lines and animal models with limited, if any, genetic diversity.

What this approach fails to consider, however, is that natural genetic variability exists between individuals in any and every population with respect to pharmaceutically-important proteins, including the protein targets of candidate drugs, the enzymes that metabolize these drugs and the proteins whose activity is modulated by such drug targets. Subtle alteration(s) in the primary nucleotide sequence of a gene encoding a pharmaceutically-important protein may be manifested as significant variation in expression, structure and/or function of the protein. Such alterations may explain the relatively high degree of uncertainty inherent in the treatment of individuals with a drug whose design is based upon a single representative example of the target or enzyme(s) involved in metabolizing the drug. For example, it is well-established that some drugs frequently have lower efficacy in some individuals than others, which means such individuals and their physicians must weigh the possible benefit of a larger dosage against a greater risk of side effects. Also, there is significant variation in how well people metabolize drugs and other exogenous chemicals, resulting in substantial interindividual variation in the toxicity and/or efficacy of such exogenous substances (Evans et al., 1999, *Science* 286:487-491). This variability in efficacy or toxicity of a drug in genetically-diverse patients makes many drugs ineffective or even dangerous in certain groups of the population, leading to the failure of such drugs in clinical trials or their early withdrawal from the market even though they could be highly beneficial for other

groups in the population. This problem significantly increases the time and cost of drug discovery and development, which is a matter of great public concern.

It is well-recognized by pharmaceutical scientists that considering the impact of the genetic variability of pharmaceutically-important proteins in the early phases of drug discovery and development is likely to reduce the failure rate of candidate and approved drugs (Marshall A 1997 *Nature Biotech* **15**:1249-52; Kleyn PW et al. 1998 *Science* **281**: 1820-21; Kola I 1999 *Curr Opin Biotech* **10**:589-92; Hill AVS et al. 1999 in *Evolution in Health and Disease* Stearns SS (Ed.) Oxford University Press, New York, pp 62-76; Meyer U.A. 1999 in *Evolution in Health and Disease* Stearns SS (Ed.) Oxford University Press, New York, pp 41-49; Kalow W et al. 1999 *Clin. Pharm. Therap.* **66**:445-7; Marshall, E 1999 *Science* **284**:406-7; Judson R et al. 2000 *Pharmacogenomics* **1**:1-12; Roses AD 2000 *Nature* **405**:857-65). However, in practice this has been difficult to do, in large part because of the time and cost required for discovering the amount of genetic variation that exists in the population (Chakravarti A 1998 *Nature Genet* **19**:216-7; Wang DG et al 1998 *Science* **280**:1077-82; Chakravarti A 1999 *Nat Genet* **21**:56-60 (suppl); Stephens JC 1999 *Mol. Diagnosis* 4:309-317; Kwok PY and Gu S 1999 *Mol. Med. Today* **5**:538-43; Davidson S 2000 *Nature Biotech* **18**:1134-5).

The standard for measuring genetic variation among individuals is the haplotype, which is the ordered combination of polymorphisms in the sequence of each form of a gene that exists in the population. Because haplotypes represent the variation across each form of a gene, they provide a more accurate and reliable measurement of genetic variation than individual polymorphisms. For example, while specific variations in gene sequences have been associated with a particular phenotype such as disease susceptibility (Roses AD *supra*; Ulbrecht M et al. 2000 *Am J Respir Crit Care Med* **161**: 469-74) and drug response (Wolfe CR et al. 2000 *BMJ* **320**:987-90; Dahl BS 1997 *Acta Psychiatr Scand* **96** (Suppl 391): 14-21), in many other cases an individual polymorphism may be found in a variety of genomic backgrounds, i.e., different haplotypes, and therefore shows no definitive coupling between the polymorphism and the causative site for the phenotype (Clark AG et al. 1998 *Am J Hum Genet* **63**:595-612; Ulbrecht M et al. 2000 *supra*; Drysdale et al. 2000 *PNAS* **97**:10483-10488). Thus, there is an unmet need in the pharmaceutical industry for information on what haplotypes exist in the population for pharmaceutically-important genes. Such haplotype information would be useful in improving the efficiency and output of several steps in the drug discovery and development process, including target validation, identifying lead compounds, and early phase clinical trials (Marshall et al., *supra*).

One pharmaceutically-important gene for the treatment of Gilbert syndrome and Crigler-Najjar syndrome is the UDP glycosyltransferase 1 (UGT1A1) gene or its encoded product. UGT1A1, also known as UGT1, is a member of the UDP-glucuronosyltransferases that are important for the conjugation and subsequent elimination of toxic xenobiotics (SWISS-PROT:P22309). UGT1A1 glucuronidates bilirubin IX-alpha to form both the IX-alpha-C8 and IX-alpha-C12 monoconjugates and diconjugate. This metabolic pathway leads to the formation of water-soluble metabolites originating from normal dietary processes, cellular catabolism, or exposure to drugs and xenobiotics (Tukey and

Strassburg, *Annu. Rev. Pharmacol. Toxicol.* 2000; 40:581-616).

Defective UGT1A1 has been implicated in both Gilbert syndrome and Crigler-Najjar syndrome. Gilbert's syndrome is shown to occur as a consequence of reduced bilirubin transferase activity. This disorder is most often detected in young adults with symptoms that are fairly nonspecific (Koiwai et al., *Hum. Mol. Genet.* 1995; 4:1183-1186). A more severe inheritable deficiency in bilirubin activity exists in Crigler-Najjar (CN). Patients with type I, which is inherited recessively, have severe hyperbilirubinemia and usually die of kernicterus, which results because of bilirubin accumulation in the nuclei of the basal ganglia and brainstem within the first year of life. Patients with type II, which is dominant, have less severe hyperbilirubinemia and usually survive into adulthood without neurologic damage. Phenobarbital, which induces the partially deficient glucuronyl transferase, can diminish the jaundice associated with this disorder (Kadakol et al., *Hum. Mutat.* 2000; 16:297-306).

The UDP glycosyltransferase 1 gene is located on chromosome 2q37 and contains 5 exons that encode a 533 amino acid protein. A reference sequence for the UGT1A1 gene is shown in Figure 1 (reverse complement of all or a portion of GenBank Accession No. AC006985.1; SEQ ID NO:1).. Reference sequences for the coding sequence (GenBankAccession No. NM_000463.1) and protein are shown in Figures 2 (SEQ ID NO:2) and 3 (SEQ ID NO:3), respectively.

One single nucleotide polymorphism in the UGT1A1 gene has been reported in the literature which corresponds to a polymorphism of guanine or adenine at nucleotide position 2826 in Figure 1. This SNP results in an amino acid variation of glycine or arginine which corresponds to amino acid position 71 in Figure 3. Maruo et al. (*Eur. J. Pediatr.* 1999; 158:547-549) identified this SNP in the UGT1A1 gene in a Japanese girl with anorexia nervosa and unconjugated hyperbilirubinemia. Akaba et al. (*Biochem. Mol. Biol. Int.* 1998; 46:21-26) reported that the gly71arg mutation of the UGT1A1 gene, which causes Gilbert syndrome, is prevalent among Japanese, Korean, and Chinese populations, with a gene frequency in those populations of 0.13, 0.23, and 0.23, respectively. Akaba et al. (*J. Hum. Genet.* 1999; 44:22-25) also showed that neonates carrying the gly71arg mutation have significantly increased bilirubin levels at days 2 to 4 and that the frequency of this mutation was significantly higher in the neonates who required phototherapy than in those who did not. These data suggest that the gly71arg mutation contributes to the high incidence of neonatal hyperbilirubinemia in Japanese.

Because of the potential for variation in the UGT1A1 gene to affect the expression and function of the encoded protein, it would be useful to know whether additional polymorphisms exist in the UGT1A1 gene, as well as how such polymorphisms are combined in different copies of the gene. Such information could be applied for studying the biological function of UGT1A1 as well as in identifying drugs targeting this protein for the treatment of disorders related to its abnormal expression or function.

SUMMARY OF THE INVENTION

Accordingly, the inventors herein have discovered 14 novel polymorphic sites in the UGT1A1 gene. These polymorphic sites (PS) correspond to the following nucleotide positions in the reverse

complement of the indicated GenBank Accession Number: 2510 (PS1), 2756 (PS2), 3155 (PS4), 3568 (PS5), 9508 (PS6), 9511 (PS7), 10091 (PS8), 10094 (PS9), 10095 (PS10), 10140 (PS11), 14423 (PS12), 14713 (PS13), 14776 (PS14) and 14971 (PS15). The polymorphisms at these sites are thymine or cytosine at PS1, cytosine or thymine at PS2, adenine or guanine at PS4, cytosine or thymine at PS5, thymine or cytosine at PS6, cytosine or thymine at PS7, cytosine or thymine at PS8, cytosine or thymine at PS9, thymine or cytosine at PS10, thymine or cytosine at PS11, adenine or thymine at PS12, cytosine or thymine at PS13, cytosine or thymine at PS14 and thymine or cytosine at PS15. In addition, the inventors have determined the identity of the alleles at these sites, as well as at the previously identified site at nucleotide position 2826 (PS3) in AC006985.2, in a human reference population of 79 unrelated individuals self-identified as belonging to one of four major population groups: African descent, Asian, Caucasian and Hispanic/Latino. From this information, the inventors deduced a set of haplotypes and haplotype pairs for PS1-15 in the UGT1A1 gene, which are shown below in Tables 4 and 3, respectively. Each of these UGT1A1 haplotypes defines a naturally-occurring isoform (also referred to herein as an "isogene") of the UGT1A1 gene that exists in the human population.

Thus, in one embodiment, the invention provides a method, composition and kit for genotyping the UGT1A1 gene in an individual. The genotyping method comprises identifying the nucleotide pair that is present at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15 in both copies of the UGT1A1 gene from the individual. A genotyping composition of the invention comprises an oligonucleotide probe or primer which is designed to specifically hybridize to a target region containing, or adjacent to, one of these novel UGT1A1 polymorphic sites. A genotyping kit of the invention comprises a set of oligonucleotides designed to genotype each of these novel UGT1A1 polymorphic sites. In a preferred embodiment, the genotyping kit comprises a set of oligonucleotides designed to genotype each of PS1-15. The genotyping method, composition, and kit are useful in determining whether an individual has one of the haplotypes in Table 4 below or has one of the haplotype pairs in Table 3 below.

The invention also provides a method for haplotyping the UGT1A1 gene in an individual. In one embodiment, the haplotyping method comprises determining, for one copy of the UGT1A1 gene, the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15. In another embodiment, the haplotyping method comprises determining whether one copy of the individual's UGT1A1 gene is defined by one of the UGT1A1 haplotypes shown in Table 4, below, or a sub-haplotype thereof. In a preferred embodiment, the haplotyping method comprises determining whether both copies of the individual's UGT1A1 gene are defined by one of the UGT1A1 haplotype pairs shown in Table 3 below, or a sub-haplotype pair thereof. The method for establishing the UGT1A1 haplotype or haplotype pair of an individual is useful for improving the efficiency and reliability of several steps in the discovery and development of drugs for treating diseases associated with UGT1A1 activity, e.g., Gilbert syndrome and Crigler-Najjar syndrome.

For example, the haplotyping method can be used by the pharmaceutical research scientist to validate UGT1A1 as a candidate target for treating a specific condition or disease predicted to be associated with UGT1A1 activity. Determining for a particular population the frequency of one or more of the individual UGT1A1 haplotypes or haplotype pairs described herein will facilitate a decision on whether to pursue UGT1A1 as a target for treating the specific disease of interest. In particular, if variable UGT1A1 activity is associated with the disease, then one or more UGT1A1 haplotypes or haplotype pairs will be found at a higher frequency in disease cohorts than in appropriately genetically matched controls. Conversely, if each of the observed UGT1A1 haplotypes are of similar frequencies in the disease and control groups, then it may be inferred that variable UGT1A1 activity has little, if any, involvement with that disease. In either case, the pharmaceutical research scientist can, without *a priori* knowledge as to the phenotypic effect of any UGT1A1 haplotype or haplotype pair, apply the information derived from detecting UGT1A1 haplotypes in an individual to decide whether modulating UGT1A1 activity would be useful in treating the disease.

The claimed invention is also useful in screening for compounds targeting UGT1A1 to treat a specific condition or disease predicted to be associated with UGT1A1 activity. For example, detecting which of the UGT1A1 haplotypes or haplotype pairs disclosed herein are present in individual members of a population with the specific disease of interest enables the pharmaceutical scientist to screen for a compound(s) that displays the highest desired agonist or antagonist activity for each of the most frequent UGT1A1 isoforms present in the disease population. Thus, without requiring any *a priori* knowledge of the phenotypic effect of any particular UGT1A1 haplotype or haplotype pair, the claimed haplotyping method provides the scientist with a tool to identify lead compounds that are more likely to show efficacy in clinical trials.

The method for haplotyping the UGT1A1 gene in an individual is also useful in the design of clinical trials of candidate drugs for treating a specific condition or disease predicted to be associated with UGT1A1 activity. For example, instead of randomly assigning patients with the disease of interest to the treatment or control group as is typically done now, determining which of the UGT1A1 haplotype(s) disclosed herein are present in individual patients enables the pharmaceutical scientist to distribute UGT1A1 haplotypes and/or haplotype pairs evenly to treatment and control groups, thereby reducing the potential for bias in the results that could be introduced by a larger frequency of a UGT1A1 haplotype or haplotype pair that had a previously unknown association with response to the drug being studied in the trial. Thus, by practicing the claimed invention, the scientist can more confidently rely on the information learned from the trial, without first determining the phenotypic effect of any UGT1A1 haplotype or haplotype pair.

In another embodiment, the invention provides a method for identifying an association between a trait and a UGT1A1 genotype, haplotype, or haplotype pair for one or more of the novel polymorphic sites described herein. The method comprises comparing the frequency of the UGT1A1 genotype, haplotype, or haplotype pair in a population exhibiting the trait with the frequency of the UGT1A1

genotype, haplotype, or haplotype pair in a reference population. A higher frequency of the UGT1A1 genotype, haplotype, or haplotype pair in the trait population than in the reference population indicates the trait is associated with the UGT1A1 genotype, haplotype, or haplotype pair. In preferred embodiments, the trait is susceptibility to a disease, severity of a disease, the staging of a disease or response to a drug. In a particularly preferred embodiment, the UGT1A1 haplotype is selected from the haplotypes shown in Table 4, or a sub-haplotype thereof. Such methods have applicability in developing diagnostic tests and therapeutic treatments for Gilbert syndrome and Crigler-Najjar syndrome.

In yet another embodiment, the invention provides an isolated polynucleotide comprising a nucleotide sequence which is a polymorphic variant of a reference sequence for the UGT1A1 gene or a fragment thereof. The reference sequence comprises SEQ ID NO:1 and the polymorphic variant comprises at least one polymorphism selected from the group consisting of cytosine at PS1, thymine at PS2, guanine at PS4, thymine at PS5, cytosine at PS6, thymine at PS7, thymine at PS8, thymine at PS9, cytosine at PS10, cytosine at PS11, thymine at PS12, thymine at PS13, thymine at PS14 and cytosine at PS15. In a preferred embodiment, the polymorphic variant comprises an additional polymorphism of adenine at PS3.

A particularly preferred polymorphic variant is an isogene of the UGT1A1 gene. A UGT1A1 isogene of the invention comprises thymine or cytosine at PS1, cytosine or thymine at PS2, guanine or adenine at PS3, adenine or guanine at PS4, cytosine or thymine at PS5, thymine or cytosine at PS6, cytosine or thymine at PS7, cytosine or thymine at PS8, cytosine or thymine at PS9, thymine or cytosine at PS10, thymine or cytosine at PS11, adenine or thymine at PS12, cytosine or thymine at PS13, cytosine or thymine at PS14 and thymine or cytosine at PS15. The invention also provides a collection of UGT1A1 isogenes, referred to herein as a UGT1A1 genome anthology.

In another embodiment, the invention provides a polynucleotide comprising a polymorphic variant of a reference sequence for a UGT1A1 cDNA or a fragment thereof. The reference sequence comprises SEQ ID NO:2 (Fig.2) and the polymorphic cDNA comprises at least one polymorphism selected from the group consisting of thymine at a position corresponding to nucleotide 141, guanine at a position corresponding to nucleotide 540, thymine at a position corresponding to nucleotide 1428 and thymine at a position corresponding to nucleotide 1491. In a preferred embodiment, the polymorphic variant comprises an additional polymorphism of adenine at a position corresponding to nucleotide 211. A particularly preferred polymorphic cDNA variant comprises the coding sequence of a UGT1A1 isogene defined by haplotypes 2- 21.

Polynucleotides complementary to these UGT1A1 genomic and cDNA variants are also provided by the invention. It is believed that polymorphic variants of the UGT1A1-gene will be useful in studying the expression and function of UGT1A1, and in expressing UGT1A1 protein for use in screening for candidate drugs to treat diseases related to UGT1A1 activity.

In other embodiments, the invention provides a recombinant expression vector comprising one

of the polymorphic genomic variants operably linked to expression regulatory elements as well as a recombinant host cell transformed or transfected with the expression vector. The recombinant vector and host cell may be used to express UGT1A1 for protein structure analysis and drug binding studies.

The present invention also provides nonhuman transgenic animals comprising one of the UGT1A1 polymorphic genomic variants described herein and methods for producing such animals. The transgenic animals are useful for studying expression of the UGT1A1 isogenes *in vivo*, for *in vivo* screening and testing of drugs targeted against UGT1A1 protein, and for testing the efficacy of therapeutic agents and compounds for Gilbert syndrome and Crigler-Najjar syndrome in a biological system.

The present invention also provides a computer system for storing and displaying polymorphism data determined for the UGT1A1 gene. The computer system comprises a computer processing unit; a display; and a database containing the polymorphism data. The polymorphism data includes the polymorphisms, the genotypes and the haplotypes identified for the UGT1A1 gene in a reference population. In a preferred embodiment, the computer system is capable of producing a display showing UGT1A1 haplotypes organized according to their evolutionary relationships.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a reference sequence for the UGT1A1 gene (Genbank Accession Number AC006985.2; contiguous lines; SEQ ID NO:1), with the start and stop positions of each region of coding sequence indicated with a bracket ([or]) and the numerical position below the sequence and the polymorphic site(s) and polymorphism(s) identified by Applicants in a reference population indicated by the variant nucleotide positioned below the polymorphic site in the sequence. SEQ ID NO:74 is equivalent to Figure 1, with the two alternative allelic variants of each polymorphic site indicated by the appropriate nucleotide symbol (R = G or A, Y = T or C, M = A or C, K = G or T, S = G or C, and W = A or T; WIPO standard ST.25).

Figure 2 illustrates a reference sequence for the UGT1A1 coding sequence (contiguous lines; SEQ ID NO:2), with the polymorphic site(s) and polymorphism(s) identified by Applicants in a reference population indicated by the variant nucleotide positioned below the polymorphic site in the sequence.

Figure 3 illustrates a reference sequence for the UGT1A1 protein (contiguous lines; SEQ ID NO:3), with the variant amino acid(s) caused by the polymorphism(s) of Figure 2 positioned below the polymorphic site in the sequence.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention is based on the discovery of novel variants of the UGT1A1 gene. As described in more detail below, the inventors herein discovered 21 isogenes of the UGT1A1 gene by characterizing the UGT1A1 gene found in genomic DNAs isolated from an Index Repository that

contains immortalized cell lines from one chimpanzee and 93 human individuals. The human individuals included a reference population of 79 unrelated individuals self-identified as belonging to one of four major population groups: Caucasian (22 individuals), African descent (20 individuals), Asian (20 individuals), or Hispanic/Latino (17 individuals). To the extent possible, the members of this reference population were organized into population subgroups by the self-identified ethnogeographic origin of their four grandparents as shown in Table 1 below.

Table 1. Population Groups in the Index Repository

Population Group	Population Subgroup	No. of Individuals
African descent		20
	Sierra Leone	1
Asian		20
	Burma	1
	China	3
	Japan	6
	Korea	1
	Philippines	5
	Vietnam	4
Caucasian		22
	British Isles	3
	British Isles/Central	4
	British Isles/Eastern	1
	Central/Eastern	1
	Eastern	3
	Central/Mediterranean	1
	Mediterranean	2
	Scandinavian	2
Hispanic/Latino		17
	Caribbean	7
	Caribbean (Spanish Descent)	2
	Central American (Spanish Descent)	1
	Mexican American	4
	South American (Spanish Descent)	3

In addition, the Index Repository contains three unrelated indigenous American Indians (one from each of North, Central and South America), one three-generation Caucasian family (from the CEPH Utah cohort) and one two-generation African-American family.

The UGT1A1 isogenes present in the human reference population are defined by haplotypes for 15 polymorphic sites in the UGT1A1 gene, 14 of which are believed to be novel. The UGT1A1 polymorphic sites identified by the inventors are referred to as PS1-15 to designate the order in which they are located in the gene (see Table 2 below), with the novel polymorphic sites referred to as PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15. Using the genotypes identified in the Index Repository for PS1-15 and the methodology described in the Examples below, the inventors herein also determined the pair of haplotypes for the UGT1A1 gene present in individual human members of this repository. The human genotypes and haplotypes found in the repository for

the UGT1A1 gene include those shown in Tables 3 and 4, respectively. The polymorphism and haplotype data disclosed herein are useful for validating whether UGT1A1 is a suitable target for drugs to treat Gilbert syndrome and Crigler-Najjar syndrome, screening for such drugs and reducing bias in clinical trials of such drugs.

In the context of this disclosure, the following terms shall be defined as follows unless otherwise indicated:

Allele - A particular form of a genetic locus, distinguished from other forms by its particular nucleotide sequence.

Candidate Gene – A gene which is hypothesized to be responsible for a disease, condition, or the response to a treatment, or to be correlated with one of these.

Gene - A segment of DNA that contains all the information for the regulated biosynthesis of an RNA product, including promoters, exons, introns, and other untranslated regions that control expression.

Genotype – An unphased 5′ to 3′ sequence of nucleotide pair(s) found at one or more polymorphic sites in a locus on a pair of homologous chromosomes in an individual. As used herein, genotype includes a full-genotype and/or a sub-genotype as described below.

Full-genotype – The unphased 5′ to 3′ sequence of nucleotide pairs found at all known polymorphic sites in a locus on a pair of homologous chromosomes in a single individual.

Sub-genotype – The unphased 5′ to 3′ sequence of nucleotides seen at a subset of the known polymorphic sites in a locus on a pair of homologous chromosomes in a single individual.

Genotyping – A process for determining a genotype of an individual.

Haplotype – A 5′ to 3′ sequence of nucleotides found at one or more polymorphic sites in a locus on a single chromosome from a single individual. As used herein, haplotype includes a full-haplotype and/or a sub-haplotype as described below.

Full-haplotype – The 5′ to 3′ sequence of nucleotides found at all known polymorphic sites in a locus on a single chromosome from a single individual.

Sub-haplotype – The 5′ to 3′ sequence of nucleotides seen at a subset of the known polymorphic sites in a locus on a single chromosome from a single individual.

Haplotype pair – The two haplotypes found for a locus in a single individual.

Haplotyping – A process for determining one or more haplotypes in an individual and includes use of family pedigrees, molecular techniques and/or statistical inference.

Haplotype data - Information concerning one or more of the following for a specific gene: a listing of the haplotype pairs in each individual in a population; a listing of the different haplotypes in a population; frequency of each haplotype in that or other populations, and any known associations between one or more haplotypes and a trait.

Isoform – A particular form of a gene, mRNA, cDNA or the protein encoded thereby, distinguished from other forms by its particular sequence and/or structure.

Isogene – One of the isoforms of a gene found in a population. An isogene contains all of the polymorphisms present in the particular isoform of the gene.

Isolated – As applied to a biological molecule such as RNA, DNA, oligonucleotide, or protein, isolated means the molecule is substantially free of other biological molecules such as nucleic acids, proteins, lipids, carbohydrates, or other material such as cellular debris and growth media. Generally, the term "isolated" is not intended to refer to a complete absence of such material or to absence of water, buffers, or salts, unless they are present in amounts that substantially interfere with the methods of the present invention.

Locus – A location on a chromosome or DNA molecule corresponding to a gene or a physical or phenotypic feature.

Naturally-occurring – A term used to designate that the object it is applied to, e.g., naturally-occurring polynucleotide or polypeptide, can be isolated from a source in nature and which has not been intentionally modified by man.

Nucleotide pair – The nucleotides found at a polymorphic site on the two copies of a chromosome from an individual.

Phased – As applied to a sequence of nucleotide pairs for two or more polymorphic sites in a locus, phased means the combination of nucleotides present at those polymorphic sites on a single copy of the locus is known.

Polymorphic site (PS) – A position within a locus at which at least two alternative sequences are found in a population, the most frequent of which has a frequency of no more than 99%.

Polymorphic variant – A gene, mRNA, cDNA, polypeptide or peptide whose nucleotide or amino acid sequence varies from a reference sequence due to the presence of a polymorphism in the gene.

Polymorphism – The sequence variation observed in an individual at a polymorphic site. Polymorphisms include nucleotide substitutions, insertions, deletions and microsatellites and may, but need not, result in detectable differences in gene expression or protein function.

Polymorphism data – Information concerning one or more of the following for a specific gene: location of polymorphic sites; sequence variation at those sites; frequency of polymorphisms in one or more populations; the different genotypes and/or haplotypes determined for the gene; frequency of one or more of these genotypes and/or haplotypes in one or more populations; any known association(s) between a trait and a genotype or a haplotype for the gene.

Polymorphism Database – A collection of polymorphism data arranged in a systematic or methodical way and capable of being individually accessed by electronic or other means.

Polynucleotide – A nucleic acid molecule comprised of single-stranded RNA or DNA or comprised of complementary, double-stranded DNA.

Population Group – A group of individuals sharing a common ethnogeographic origin.

Reference Population – A group of subjects or individuals who are predicted to be

representative of the genetic variation found in the general population. Typically, the reference population represents the genetic variation in the population at a certainty level of at least 85%, preferably at least 90%, more preferably at least 95% and even more preferably at least 99%.

Single Nucleotide Polymorphism (SNP) – Typically, the specific pair of nucleotides observed at a single polymorphic site. In rare cases, three or four nucleotides may be found.

Subject – A human individual whose genotypes or haplotypes or response to treatment or disease state are to be determined.

Treatment - A stimulus administered internally or externally to a subject.

Unphased – As applied to a sequence of nucleotide pairs for two or more polymorphic sites in a locus, unphased means the combination of nucleotides present at those polymorphic sites on a single copy of the locus is not known.

As discussed above, information on the identity of genotypes and haplotypes for the UGT1A1 gene of any particular individual as well as the frequency of such genotypes and haplotypes in any particular population of individuals is expected to be useful for a variety of drug discovery and development applications. Thus, the invention also provides compositions and methods for detecting the novel UGT1A1 polymorphisms and haplotypes identified herein.

The compositions comprise at least one UGT1A1 genotyping oligonucleotide. In one embodiment, a UGT1A1 genotyping oligonucleotide is a probe or primer capable of hybridizing to a target region that is located close to, or that contains, one of the novel polymorphic sites described herein. As used herein, the term "oligonucleotide" refers to a polynucleotide molecule having less than about 100 nucleotides. A preferred oligonucleotide of the invention is 10 to 35 nucleotides long. More preferably, the oligonucleotide is between 15 and 30, and most preferably, between 20 and 25 nucleotides in length. The exact length of the oligonucleotide will depend on many factors that are routinely considered and practiced by the skilled artisan. The oligonucleotide may be comprised of any phosphorylation state of ribonucleotides, deoxyribonucleotides, and acyclic nucleotide derivatives, and other functionally equivalent derivatives. Alternatively, oligonucleotides may have a phosphate-free backbone, which may be comprised of linkages such as carboxymethyl, acetamidate, carbamate, polyamide (peptide nucleic acid (PNA)) and the like (Varma, R. in Molecular Biology and Biotechnology, A Comprehensive Desk Reference, Ed. R. Meyers, VCH Publishers, Inc. (1995), pages 617-620). Oligonucleotides of the invention may be prepared by chemical synthesis using any suitable methodology known in the art, or may be derived from a biological sample, for example, by restriction digestion. The oligonucleotides may be labeled, according to any technique known in the art, including use of radiolabels, fluorescent labels, enzymatic labels, proteins, haptens, antibodies, sequence tags and the like.

Genotyping oligonucleotides of the invention must be capable of specifically hybridizing to a target region of a UGT1A1 polynucleotide, i.e., a UGT1A1 isogene. As used herein, specific hybridization means the oligonucleotide forms an anti-parallel double-stranded structure with the target

region under certain hybridizing conditions, while failing to form such a structure when incubated with a non-target region or a non-UGT1A1 polynucleotide under the same hybridizing conditions. Preferably, the oligonucleotide specifically hybridizes to the target region under conventional high stringency conditions. The skilled artisan can readily design and test oligonucleotide probes and primers suitable for detecting polymorphisms in the UGT1A1 gene using the polymorphism information provided herein in conjunction with the known sequence information for the UGT1A1 gene and routine techniques.

A nucleic acid molecule such as an oligonucleotide or polynucleotide is said to be a "perfect" or "complete" complement of another nucleic acid molecule if every nucleotide of one of the molecules is complementary to the nucleotide at the corresponding position of the other molecule. A nucleic acid molecule is "substantially complementary" to another molecule if it hybridizes to that molecule with sufficient stability to remain in a duplex form under conventional low-stringency conditions. Conventional hybridization conditions are described, for example, by Sambrook J. et al., in Molecular Cloning, A Laboratory Manual, 2nd Edition, Cold Spring Harbor Press, Cold Spring Harbor, NY (1989) and by Haymes, B.D. et al. in Nucleic Acid Hybridization, A Practical Approach, IRL Press, Washington, D.C. (1985). While perfectly complementary oligonucleotides are preferred for detecting polymorphisms, departures from complete complementarity are contemplated where such departures do not prevent the molecule from specifically hybridizing to the target region. For example, an oligonucleotide primer may have a non-complementary fragment at its 5′ end, with the remainder of the primer being complementary to the target region. Alternatively, non-complementary nucleotides may be interspersed into the oligonucleotide probe or primer as long as the resulting probe or primer is still capable of specifically hybridizing to the target region.

Preferred genotyping oligonucleotides of the invention are allele-specific oligonucleotides. As used herein, the term allele-specific oligonucleotide (ASO) means an oligonucleotide that is able, under sufficiently stringent conditions, to hybridize specifically to one allele of a gene, or other locus, at a target region containing a polymorphic site while not hybridizing to the corresponding region in another allele(s). As understood by the skilled artisan, allele-specificity will depend upon a variety of readily optimized stringency conditions, including salt and formamide concentrations, as well as temperatures for both the hybridization and washing steps. Examples of hybridization and washing conditions typically used for ASO probes are found in Kogan et al., "Genetic Prediction of Hemophilia A" in PCR Protocols, A Guide to Methods and Applications, Academic Press, 1990 and Ruaño et al., 87 *Proc. Natl. Acad. Sci. USA* 6296-6300, 1990. Typically, an ASO will be perfectly complementary to one allele while containing a single mismatch for another allele.

Allele-specific oligonucleotides of the invention include ASO probes and ASO primers. ASO probes which usually provide good discrimination between different alleles are those in which a central position of the oligonucleotide probe aligns with the polymorphic site in the target region (e.g., approximately the 7th or 8th position in a 15mer, the 8th or 9th position in a 16mer, and the 10th or 11th

position in a 20mer). An ASO primer of the invention has a 3′ terminal nucleotide, or preferably a 3′ penultimate nucleotide, that is complementary to only one nucleotide of a particular SNP, thereby acting as a primer for polymerase-mediated extension only if the allele containing that nucleotide is present. ASO probes and primers hybridizing to either the coding or noncoding strand are contemplated by the invention.

ASO probes and primers listed below use the appropriate nucleotide symbol (R= G or A, Y= T or C, M= A or C, K= G or T, S= G or C, and W= A or T; WIPO standard ST.25) at the position of the polymorphic site to represent the two alternative allelic variants observed at that polymorphic site.

A preferred ASO probe for detecting UGT1A1 gene polymorphisms comprises a nucleotide sequence, listed 5′ to 3′, selected from the group consisting of:

```
CTTTTTAYAGTCACG   (SEQ ID NO:4) and its complement,
GGGCCATYCAGCAGC   (SEQ ID NO:5) and its complement,
GCCTGGARTTTGAGG   (SEQ ID NO:6) and its complement,
TGCTGAGYAAGCATT   (SEQ ID NO:7) and its complement,
GATTCTAYACCATGG   (SEQ ID NO:8) and its complement,
TCTATACYATGGCCT   (SEQ ID NO:9) and its complement,
CAGAGGAYCCCTGTT   (SEQ ID NO:10) and its complement,
AGGACCCYTGTTTTC   (SEQ ID NO:11) and its complement,
GGACCCCYGTTTTCT   (SEQ ID NO:12) and its complement,
TATTATGYTCTTTCT   (SEQ ID NO:13) and its complement,
GGGCAACWGGGCAAG   (SEQ ID NO:14) and its complement,
TGCGCCCYGCAGCCC   (SEQ ID NO:15) and its complement,
TCTTGGCYGTCGTGC   (SEQ ID NO:16) and its complement, and
AATTCATYTTATTCT   (SEQ ID NO:17) and its complement.
```

A preferred ASO primer for detecting UGT1A1 gene polymorphisms comprises a nucleotide sequence, listed 5′ to 3′, selected from the group consisting of:

```
TGACAGCTTTTTAYA (SEQ ID NO:18); GTGTCACGTGACTRT (SEQ ID NO:19);
TGCTTGGGGCCATYC (SEQ ID NO:20); GCTGCAGCTGCTGRA (SEQ ID NO:21);
CATGCAGCCTGGART (SEQ ID NO:22); GGGTAGCCTCAAAYT (SEQ ID NO:23);
GATATATGCTGAGYA (SEQ ID NO:24); TCTCAGAATGCTTRC (SEQ ID NO:25);
TAAGAAGATTCTAYA (SEQ ID NO:26); ATGAGGCCATGGTRT (SEQ ID NO:27);
GAAGATTCTATACYA (SEQ ID NO:28); GATATGAGGCCATRG (SEQ ID NO:29);
TTCCTTCAGAGGAYC (SEQ ID NO:30); CTAGAAAACAGGGRT (SEQ ID NO:31);
CTTCAGAGGACCCYT (SEQ ID NO:32); TAACTAGAAAACARG (SEQ ID NO:33);
TTCAGAGGACCCCYG (SEQ ID NO:34); CTAACTAGAAAACRG (SEQ ID NO:35);
TAATCATATTATGYT (SEQ ID NO:36); ACGTAAAGAAAGARC (SEQ ID NO:37);
CAGCCCGGGCAACWG (SEQ ID NO:38); CAGAGTCTTGCCCWG (SEQ ID NO:39);
CACACCTGCGCCCYG (SEQ ID NO:40); GGTCGTGGGCTGCRG (SEQ ID NO:41);
GTTTCCTCTTGGCYG (SEQ ID NO:42); CTGTCAGCACGACRG (SEQ ID NO:43);
GTGTTAAATTCATYT (SEQ ID NO:44); and TTAATAAGAATAARA (SEQ ID NO:45).
```

Other genotyping oligonucleotides of the invention hybridize to a target region located one to several nucleotides downstream of one of the novel polymorphic sites identified herein. Such oligonucleotides are useful in polymerase-mediated primer extension methods for detecting one of the novel polymorphisms described herein and therefore such genotyping oligonucleotides are referred to herein as “primer-extension oligonucleotides”. In a preferred embodiment, the 3′-terminus of a primer-

extension oligonucleotide is a deoxynucleotide complementary to the nucleotide located immediately adjacent to the polymorphic site.

A particularly preferred oligonucleotide primer for detecting UGT1A1 gene polymorphisms by primer extension terminates in a nucleotide sequence, listed 5′ to 3′, selected from the group consisting of:

```
CAGCTTTTTA  (SEQ ID NO:46); TCACGTGACT  (SEQ ID NO:47);
TTGGGGCCAT  (SEQ ID NO:48); GCAGCTGCTG  (SEQ ID NO:49);
GCAGCCTGGA  (SEQ ID NO:50); TAGCCTCAAA  (SEQ ID NO:51);
ATATGCTGAG  (SEQ ID NO:52); CAGAATGCTT  (SEQ ID NO:53);
GAAGATTCTA  (SEQ ID NO:54); AGGCCATGGT  (SEQ ID NO:55);
GATTCTATAC  (SEQ ID NO:56); ATGAGGCCAT  (SEQ ID NO:57);
CTTCAGAGGA  (SEQ ID NO:58); GAAAACAGGG  (SEQ ID NO:59);
CAGAGGACCC  (SEQ ID NO:60); CTAGAAAACA  (SEQ ID NO:61);
AGAGGACCCC  (SEQ ID NO:62); ACTAGAAAAC  (SEQ ID NO:63);
TCATATTATG  (SEQ ID NO:64); TAAAGAAAGA  (SEQ ID NO:65);
CCCGGGCAAC  (SEQ ID NO:66); AGTCTTGCCC  (SEQ ID NO:67);
ACCTGCGCCC  (SEQ ID NO:68); CGTGGGCTGC  (SEQ ID NO:69);
TCCTCTTGGC  (SEQ ID NO:70); TCAGCACGAC  (SEQ ID NO:71);
TTAAATTCAT  (SEQ ID NO:72); and ATAAGAATAA  (SEQ ID NO:73).
```

In some embodiments, a composition contains two or more differently labeled genotyping oligonucleotides for simultaneously probing the identity of nucleotides at two or more polymorphic sites. It is also contemplated that primer compositions may contain two or more sets of allele-specific primer pairs to allow simultaneous targeting and amplification of two or more regions containing a polymorphic site.

UGT1A1 genotyping oligonucleotides of the invention may also be immobilized on or synthesized on a solid surface such as a microchip, bead, or glass slide (see, e.g., WO 98/20020 and WO 98/20019). Such immobilized genotyping oligonucleotides may be used in a variety of polymorphism detection assays, including but not limited to probe hybridization and polymerase extension assays. Immobilized UGT1A1 genotyping oligonucleotides of the invention may comprise an ordered array of oligonucleotides designed to rapidly screen a DNA sample for polymorphisms in multiple genes at the same time.

In another embodiment, the invention provides a kit comprising at least two genotyping oligonucleotides packaged in separate containers. The kit may also contain other components such as hybridization buffer (where the oligonucleotides are to be used as a probe) packaged in a separate container. Alternatively, where the oligonucleotides are to be used to amplify a target region, the kit may contain, packaged in separate containers, a polymerase and a reaction buffer optimized for primer extension mediated by the polymerase, such as PCR.

The above described oligonucleotide compositions and kits are useful in methods for genotyping and/or haplotyping the UGT1A1 gene in an individual. As used herein, the terms "UGT1A1 genotype" and "UGT1A1 haplotype" mean the genotype or haplotype contains the nucleotide pair or nucleotide, respectively, that is present at one or more of the novel polymorphic sites

described herein and may optionally also include the nucleotide pair or nucleotide present at one or more additional polymorphic sites in the UGT1A1 gene. The additional polymorphic sites may be currently known polymorphic sites or sites that are subsequently discovered.

One embodiment of the genotyping method involves isolating from the individual a nucleic acid sample comprising the two copies of the UGT1A1 gene, or a fragment thereof, that are present in the individual, and determining the identity of the nucleotide pair at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15 in the two copies to assign a UGT1A1 genotype to the individual. As will be readily understood by the skilled artisan, the two "copies" of a gene in an individual may be the same allele or may be different alleles. In a preferred embodiment of the genotyping method, the identity of the nucleotide pair at PS3 is also determined. In a particularly preferred embodiment, the genotyping method comprises determining the identity of the nucleotide pair at each of PS1-15.

Typically, the nucleic acid sample is isolated from a biological sample taken from the individual, such as a blood sample or tissue sample. Suitable tissue samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair. The nucleic acid sample may be comprised of genomic DNA, mRNA, or cDNA and, in the latter two cases, the biological sample must be obtained from a tissue in which the UGT1A1 gene is expressed. Furthermore it will be understood by the skilled artisan that mRNA or cDNA preparations would not be used to detect polymorphisms located in introns or in 5′ and 3′ untranslated regions. If a UGT1A1 gene fragment is isolated, it must contain the polymorphic site(s) to be genotyped.

One embodiment of the haplotyping method comprises isolating from the individual a nucleic acid sample containing only one of the two copies of the UGT1A1 gene, or a fragment thereof, that is present in the individual and determining in that copy the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15 in that copy to assign a UGT1A1 haplotype to the individual. The nucleic acid may be isolated using any method capable of separating the two copies of the UGT1A1 gene or fragment such as one of the methods described above for preparing UGT1A1 isogenes, with targeted *in vivo* cloning being the preferred approach. As will be readily appreciated by those skilled in the art, any individual clone will only provide haplotype information on one of the two UGT1A1 gene copies present in an individual. If haplotype information is desired for the individual's other copy, additional UGT1A1 clones will need to be examined. Typically, at least five clones should be examined to have more than a 90% probability of haplotyping both copies of the UGT1A1 gene in an individual. In some embodiments, the haplotyping method also comprises identifying the nucleotide at PS3. In a particularly preferred embodiment, the nucleotide at each of PS1-15 is identified.

In another embodiment, the haplotyping method comprises determining whether an individual has one or more of the UGT1A1 haplotypes shown in Table 4. This can be accomplished by identifying, for one or both copies of the individual's UGT1A1 gene, the phased sequence of

nucleotides present at each of PS1-15. The present invention also contemplates that typically only a subset of PS1-15 will need to be directly examined to assign to an individual one or more of the haplotypes shown in Table 4. This is because at least one polymorphic site in a gene is frequently in strong linkage disequilibrium with one or more other polymorphic sites in that gene (Drysdale, CM et al. 2000 *PNAS* **97**:10483-10488; Rieder MJ et al. 1999 *Nature Genetics* **22**:59-62). Two sites are said to be in linkage disequilibrium if the presence of a particular variant at one site enhances the predictability of another variant at the second site (Stephens, JC 1999, *Mol. Diag.* **4**:309-317). Techniques for determining whether any two polymorphic sites are in linkage disequilibrium are well-known in the art (Weir B.S. 1996 *Genetic Data Analysis II*, Sinauer Associates, Inc. Publishers, Sunderland, MA).

In a preferred embodiment, a UGT1A1 haplotype pair is determined for an individual by identifying the phased sequence of nucleotides at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15 in each copy of the UGT1A1 gene that is present in the individual. In a particularly preferred embodiment, the haplotyping method comprises identifying the phased sequence of nucleotides at each of PS1-15 in each copy of the UGT1A1 gene. When haplotyping both copies of the gene, the identifying step is preferably performed with each copy of the gene being placed in separate containers. However, it is also envisioned that if the two copies are labeled with different tags, or are otherwise separately distinguishable or identifiable, it could be possible in some cases to perform the method in the same container. For example, if first and second copies of the gene are labeled with different first and second fluorescent dyes, respectively, and an allele-specific oligonucleotide labeled with yet a third different fluorescent dye is used to assay the polymorphic site(s), then detecting a combination of the first and third dyes would identify the polymorphism in the first gene copy while detecting a combination of the second and third dyes would identify the polymorphism in the second gene copy.

In both the genotyping and haplotyping methods, the identity of a nucleotide (or nucleotide pair) at a polymorphic site(s) may be determined by amplifying a target region(s) containing the polymorphic site(s) directly from one or both copies of the UGT1A1 gene, or a fragment thereof, and the sequence of the amplified region(s) determined by conventional methods. It will be readily appreciated by the skilled artisan that only one nucleotide will be detected at a polymorphic site in individuals who are homozygous at that site, while two different nucleotides will be detected if the individual is heterozygous for that site. The polymorphism may be identified directly, known as positive-type identification, or by inference, referred to as negative-type identification. For example, where a SNP is known to be guanine and cytosine in a reference population, a site may be positively determined to be either guanine or cytosine for an individual homozygous at that site, or both guanine and cytosine, if the individual is heterozygous at that site. Alternatively, the site may be negatively determined to be not guanine (and thus cytosine/cytosine) or not cytosine (and thus guanine/guanine).

The target region(s) may be amplified using any oligonucleotide-directed amplification method,

including but not limited to polymerase chain reaction (PCR) (U.S. Patent No. 4,965,188), ligase chain reaction (LCR) (Barany et al., *Proc. Natl. Acad. Sci. USA* 88:189-193, 1991; WO90/01069), and oligonucleotide ligation assay (OLA) (Landegren et al., *Science* 241:1077-1080, 1988).

Other known nucleic acid amplification procedures may be used to amplify the target region including transcription-based amplification systems (U.S. Patent No. 5,130,238; EP 329,822; U.S. Patent No. 5,169,766, WO89/06700) and isothermal methods (Walker et al., *Proc. Natl. Acad. Sci. USA* **89**:392-396, 1992).

A polymorphism in the target region may also be assayed before or after amplification using one of several hybridization-based methods known in the art. Typically, allele-specific oligonucleotides are utilized in performing such methods. The allele-specific oligonucleotides may be used as differently labeled probe pairs, with one member of the pair showing a perfect match to one variant of a target sequence and the other member showing a perfect match to a different variant. In some embodiments, more than one polymorphic site may be detected at once using a set of allele-specific oligonucleotides or oligonucleotide pairs. Preferably, the members of the set have melting temperatures within 5°C, and more preferably within 2°C, of each other when hybridizing to each of the polymorphic sites being detected.

Hybridization of an allele-specific oligonucleotide to a target polynucleotide may be performed with both entities in solution, or such hybridization may be performed when either the oligonucleotide or the target polynucleotide is covalently or noncovalently affixed to a solid support. Attachment may be mediated, for example, by antibody-antigen interactions, poly-L-Lys, streptavidin or avidin-biotin, salt bridges, hydrophobic interactions, chemical linkages, UV cross-linking baking, etc. Allele-specific oligonucleotides may be synthesized directly on the solid support or attached to the solid support subsequent to synthesis. Solid-supports suitable for use in detection methods of the invention include substrates made of silicon, glass, plastic, paper and the like, which may be formed, for example, into wells (as in 96-well plates), slides, sheets, membranes, fibers, chips, dishes, and beads. The solid support may be treated, coated or derivatized to facilitate the immobilization of the allele-specific oligonucleotide or target nucleic acid.

The genotype or haplotype for the UGT1A1 gene of an individual may also be determined by hybridization of a nucleic acid sample containing one or both copies of the gene, or fragment(s) thereof, to nucleic acid arrays and subarrays such as described in WO 95/11995. The arrays would contain a battery of allele-specific oligonucleotides representing each of the polymorphic sites to be included in the genotype or haplotype.

The identity of polymorphisms may also be determined using a mismatch detection technique, including but not limited to the RNase protection method using riboprobes (Winter et al., *Proc. Natl. Acad. Sci. USA* 82:7575, 1985; Meyers et al., *Science* 230:1242, 1985) and proteins which recognize nucleotide mismatches, such as the E. coli mutS protein (Modrich, P. *Ann. Rev. Genet.* 25:229-253, 1991). Alternatively, variant alleles can be identified by single strand conformation polymorphism

(SSCP) analysis (Orita et al., *Genomics* 5:874-879, 1989; Humphries et al., in Molecular Diagnosis of Genetic Diseases, R. Elles, ed., pp. 321-340, 1996) or denaturing gradient gel electrophoresis (DGGE) (Wartell et al., *Nucl. Acids Res.* 18:2699-2706, 1990; Sheffield et al., *Proc. Natl. Acad. Sci. USA* 86:232-236, 1989).

A polymerase-mediated primer extension method may also be used to identify the polymorphism(s). Several such methods have been described in the patent and scientific literature and include the "Genetic Bit Analysis" method (WO92/15712) and the ligase/polymerase mediated genetic bit analysis (U.S. Patent 5,679,524. Related methods are disclosed in WO91/02087, WO90/09455, WO95/17676, U.S. Patent Nos. 5,302,509, and 5,945,283. Extended primers containing a polymorphism may be detected by mass spectrometry as described in U.S. Patent No. 5,605,798. Another primer extension method is allele-specific PCR (Ruaño et al., *Nucl. Acids Res.* 17:8392, 1989; Ruaño et al., *Nucl. Acids Res.* 19, 6877-6882, 1991; WO 93/22456; Turki et al., *J. Clin. Invest.* 95:1635-1641, 1995). In addition, multiple polymorphic sites may be investigated by simultaneously amplifying multiple regions of the nucleic acid using sets of allele-specific primers as described in Wallace et al. (WO89/10414).

In addition, the identity of the allele(s) present at any of the novel polymorphic sites described herein may be indirectly determined by genotyping another polymorphic site that is in linkage disequilibrium with the polymorphic site that is of interest. Polymorphic sites in linkage disequilibrium with the presently disclosed polymorphic sites may be located in regions of the gene or in other genomic regions not examined herein. Genotyping of a polymorphic site in linkage disequilibrium with the novel polymorphic sites described herein may be performed by, but is not limited to, any of the above-mentioned methods for detecting the identity of the allele at a polymorphic site.

In another aspect of the invention, an individual's UGT1A1 haplotype pair is predicted from its UGT1A1 genotype using information on haplotype pairs known to exist in a reference population. In its broadest embodiment, the haplotyping prediction method comprises identifying a UGT1A1 genotype for the individual at two or more UGT1A1 polymorphic sites described herein, enumerating all possible haplotype pairs which are consistent with the genotype, accessing data containing UGT1A1 haplotype pairs identified in a reference population, and assigning a haplotype pair to the individual that is consistent with the data. In one embodiment, the reference haplotype pairs include the UGT1A1 haplotype pairs shown in Table 3.

Generally, the reference population should be composed of randomly-selected individuals representing the major ethnogeographic groups of the world. A preferred reference population for use in the methods of the present invention comprises an approximately equal number of individuals from Caucasian, African American, Asian and Hispanic-Latino population groups with the minimum number of each group being chosen based on how rare a haplotype one wants to be guaranteed to see. For example, if one wants to have a q% chance of not missing a haplotype that exists in the population at a p% frequency of occurring in the reference population, the number of individuals (n) who must be

sampled is given by 2n=log(1-q)/log(1-p) where p and q are expressed as fractions. A preferred reference population allows the detection of any haplotype whose frequency is at least 10% with about 99% certainty and comprises about 20 unrelated individuals from each of the four population groups named above. A particularly preferred reference population includes a 3-generation family representing one or more of the four population groups to serve as controls for checking quality of haplotyping procedures.

In a preferred embodiment, the haplotype frequency data for each ethnogeographic group is examined to determine whether it is consistent with Hardy-Weinberg equilibrium. Hardy-Weinberg equilibrium (D.L. Hartl et al., Principles of Population Genomics, Sinauer Associates (Sunderland, MA), 3rd Ed., 1997) postulates that the frequency of finding the haplotype pair H_1 / H_2 is equal to $p_{H\text{-}W}(H_1 / H_2) = 2p(H_1)p(H_2)$ if $H_1 \neq H_2$ and $p_{H\text{-}W}(H_1 / H_2) = p(H_1)p(H_2)$ if $H_1 = H_2$. A statistically significant difference between the observed and expected haplotype frequencies could be due to one or more factors including significant inbreeding in the population group, strong selective pressure on the gene, sampling bias, and/or errors in the genotyping process. If large deviations from Hardy-Weinberg equilibrium are observed in an ethnogeographic group, the number of individuals in that group can be increased to see if the deviation is due to a sampling bias. If a larger sample size does not reduce the difference between observed and expected haplotype pair frequencies, then one may wish to consider haplotyping the individual using a direct haplotyping method such as, for example, CLASPER System™ technology (U.S. Patent No. 5,866,404), single molecule dilution, or allele-specific long-range PCR (Michalotos-Beloin et al., *Nucleic Acids Res.* 24:4841-4843, 1996).

In one embodiment of this method for predicting a UGT1A1 haplotype pair for an individual, the assigning step involves performing the following analysis. First, each of the possible haplotype pairs is compared to the haplotype pairs in the reference population. Generally, only one of the haplotype pairs in the reference population matches a possible haplotype pair and that pair is assigned to the individual. Occasionally, only one haplotype represented in the reference haplotype pairs is consistent with a possible haplotype pair for an individual, and in such cases the individual is assigned a haplotype pair containing this known haplotype and a new haplotype derived by subtracting the known haplotype from the possible haplotype pair. Alternatively, the haplotype pair in an individual may be predicted from the individual's genotype for that gene using reported methods (e.g., Clark et al. 1990 *Mol Bio Evol* 7:111-22) or through a commercial haplotyping service such as offered by Genaissance Pharmaceuticals, Inc. (New Haven, CT). In rare cases, either no haplotypes in the reference population are consistent with the possible haplotype pairs, or alternatively, multiple reference haplotype pairs are consistent with the possible haplotype pairs. In such cases, the individual is preferably haplotyped using a direct molecular haplotyping method such as, for example, CLASPER System™ technology (U.S. Patent No. 5,866,404), SMD, or allele-specific long-range PCR (Michalotos-Beloin et al., *supra*).

The invention also provides a method for determining the frequency of a UGT1A1 genotype,

haplotype, or haplotype pair in a population. The method comprises, for each member of the population, determining the genotype or the haplotype pair for the novel UGT1A1 polymorphic sites described herein, and calculating the frequency any particular genotype, haplotype, or haplotype pair is found in the population. The population may be a reference population, a family population, a same sex population, a population group, or a trait population (e.g., a group of individuals exhibiting a trait of interest such as a medical condition or response to a therapeutic treatment).

In another aspect of the invention, frequency data for UGT1A1 genotypes, haplotypes, and/or haplotype pairs are determined in a reference population and used in a method for identifying an association between a trait and a UGT1A1 genotype, haplotype, or haplotype pair. The trait may be any detectable phenotype, including but not limited to susceptibility to a disease or response to a treatment. The method involves obtaining data on the frequency of the genotype(s), haplotype(s), or haplotype pair(s) of interest in a reference population as well as in a population exhibiting the trait. Frequency data for one or both of the reference and trait populations may be obtained by genotyping or haplotyping each individual in the populations using one of the methods described above. The haplotypes for the trait population may be determined directly or, alternatively, by the predictive genotype to haplotype approach described above. In another embodiment, the frequency data for the reference and/or trait populations is obtained by accessing previously determined frequency data, which may be in written or electronic form. For example, the frequency data may be present in a database that is accessible by a computer. Once the frequency data is obtained, the frequencies of the genotype(s), haplotype(s), or haplotype pair(s) of interest in the reference and trait populations are compared. In a preferred embodiment, the frequencies of all genotypes, haplotypes, and/or haplotype pairs observed in the populations are compared. If a particular UGT1A1 genotype, haplotype, or haplotype pair is more frequent in the trait population than in the reference population at a statistically significant amount, then the trait is predicted to be associated with that UGT1A1 genotype, haplotype, or haplotype pair. Preferably, the UGT1A1 genotype, haplotype, or haplotype pair being compared in the trait and reference populations is selected from the full-genotypes and full-haplotypes shown in Tables 3 and 4, or from sub-genotypes and sub-haplotypes derived from these genotypes and haplotypes.

In a preferred embodiment of the method, the trait of interest is a clinical response exhibited by a patient to some therapeutic treatment, for example, response to a drug targeting UGT1A1 or response to a therapeutic treatment for a medical condition. As used herein, "medical condition" includes but is not limited to any condition or disease manifested as one or more physical and/or psychological symptoms for which treatment is desirable, and includes previously and newly identified diseases and other disorders. As used herein the term "clinical response" means any or all of the following: a quantitative measure of the response, no response, and adverse response (i.e., side effects).

In order to deduce a correlation between clinical response to a treatment and a UGT1A1 genotype, haplotype, or haplotype pair, it is necessary to obtain data on the clinical responses exhibited by a population of individuals who received the treatment, hereinafter the "clinical population". This

clinical data may be obtained by analyzing the results of a clinical trial that has already been run and/or the clinical data may be obtained by designing and carrying out one or more new clinical trials. As used herein, the term "clinical trial" means any research study designed to collect clinical data on responses to a particular treatment, and includes but is not limited to phase I, phase II and phase III clinical trials. Standard methods are used to define the patient population and to enroll subjects.

It is preferred that the individuals included in the clinical population have been graded for the existence of the medical condition of interest. This is important in cases where the symptom(s) being presented by the patients can be caused by more than one underlying condition, and where treatment of the underlying conditions are not the same. An example of this would be where patients experience breathing difficulties that are due to either asthma or respiratory infections. If both sets were treated with an asthma medication, there would be a spurious group of apparent non-responders that did not actually have asthma. These people would affect the ability to detect any correlation between haplotype and treatment outcome. This grading of potential patients could employ a standard physical exam or one or more lab tests. Alternatively, grading of patients could use haplotyping for situations where there is a strong correlation between haplotype pair and disease susceptibility or severity.

The therapeutic treatment of interest is administered to each individual in the trial population and each individual's response to the treatment is measured using one or more predetermined criteria. It is contemplated that in many cases, the trial population will exhibit a range of responses and that the investigator will choose the number of responder groups (e.g., low, medium, high) made up by the various responses. In addition, the UGT1A1 gene for each individual in the trial population is genotyped and/or haplotyped, which may be done before or after administering the treatment.

After both the clinical and polymorphism data have been obtained, correlations between individual response and UGT1A1 genotype or haplotype content are created. Correlations may be produced in several ways. In one method, individuals are grouped by their UGT1A1 genotype or haplotype (or haplotype pair) (also referred to as a polymorphism group), and then the averages and standard deviations of clinical responses exhibited by the members of each polymorphism group are calculated.

These results are then analyzed to determine if any observed variation in clinical response between polymorphism groups is statistically significant. Statistical analysis methods which may be used are described in L.D. Fisher and G. vanBelle, "Biostatistics: A Methodology for the Health Sciences", Wiley-Interscience (New York) 1993. This analysis may also include a regression calculation of which polymorphic sites in the UGT1A1 gene give the most significant contribution to the differences in phenotype. One regression model useful in the invention is described in PCT Application Serial No. PCT/US00/17540, entitled "Methods for Obtaining and Using Haplotype Data".

A second method for finding correlations between UGT1A1 haplotype content and clinical responses uses predictive models based on error-minimizing optimization algorithms. One of many possible optimization algorithms is a genetic algorithm (R. Judson, "Genetic Algorithms and Their Uses

in Chemistry" in Reviews in Computational Chemistry, Vol. 10, pp. 1-73, K. B. Lipkowitz and D. B. Boyd, eds. (VCH Publishers, New York, 1997). Simulated annealing (Press et al., "Numerical Recipes in C: The Art of Scientific Computing", Cambridge University Press (Cambridge) 1992, Ch. 10), neural networks (E. Rich and K. Knight, "Artificial Intelligence", 2nd Edition (McGraw-Hill, New York, 1991, Ch. 18), standard gradient descent methods (Press et al., *supra,* Ch. 10), or other global or local optimization approaches (see discussion in Judson, *supra*) could also be used. Preferably, the correlation is found using a genetic algorithm approach as described in PCT Application Serial No. PCT/US00/17540.

Correlations may also be analyzed using analysis of variation (ANOVA) techniques to determine how much of the variation in the clinical data is explained by different subsets of the polymorphic sites in the UGT1A1 gene. As described in PCT Application Serial No. PCT/US00/17540, ANOVA is used to test hypotheses about whether a response variable is caused by or correlated with one or more traits or variables that can be measured (Fisher and vanBelle, *supra*, Ch. 10).

From the analyses described above, a mathematical model may be readily constructed by the skilled artisan that predicts clinical response as a function of UGT1A1 genotype or haplotype content. Preferably, the model is validated in one or more follow-up clinical trials designed to test the model.

The identification of an association between a clinical response and a genotype or haplotype (or haplotype pair) for the UGT1A1 gene may be the basis for designing a diagnostic method to determine those individuals who will or will not respond to the treatment, or alternatively, will respond at a lower level and thus may require more treatment, i.e., a greater dose of a drug. The diagnostic method may take one of several forms: for example, a direct DNA test (i.e., genotyping or haplotyping one or more of the polymorphic sites in the UGT1A1 gene), a serological test, or a physical exam measurement. The only requirement is that there be a good correlation between the diagnostic test results and the underlying UGT1A1 genotype or haplotype that is in turn correlated with the clinical response. In a preferred embodiment, this diagnostic method uses the predictive haplotyping method described above.

In another embodiment, the invention provides an isolated polynucleotide comprising a polymorphic variant of the UGT1A1 gene or a fragment of the gene which contains at least one of the novel polymorphic sites described herein. The nucleotide sequence of a variant UGT1A1 gene is identical to the reference genomic sequence for those portions of the gene examined, as described in the Examples below, except that it comprises a different nucleotide at one or more of the novel polymorphic sites PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15, and may also comprise an additional polymorphism of adenine at PS3. Similarly, the nucleotide sequence of a variant fragment of the UGT1A1 gene is identical to the corresponding portion of the reference sequence except for having a different nucleotide at one or more of the novel polymorphic sites described herein. Thus, the invention specifically does not include polynucleotides comprising a nucleotide sequence identical to the reference sequence of the UGT1A1 gene, which is defined by

haplotype 1, (or other reported UGT1A1 sequences) or to portions of the reference sequence (or other reported UGT1A1 sequences), except for genotyping oligonucleotides as described below.

The location of a polymorphism in a variant gene or fragment is identified by aligning its sequence against SEQ ID NO:1. The polymorphism is selected from the group consisting of cytosine at PS1, thymine at PS2, guanine at PS4, thymine at PS5, cytosine at PS6, thymine at PS7, thymine at PS8, thymine at PS9, cytosine at PS10, cytosine at PS11, thymine at PS12, thymine at PS13, thymine at PS14 and cytosine at PS15. In a preferred embodiment, the polymorphic variant comprises a naturally-occurring isogene of the UGT1A1 gene which is defined by any one of haplotypes 2- 21 shown in Table 4 below.

Polymorphic variants of the invention may be prepared by isolating a clone containing the UGT1A1 gene from a human genomic library. The clone may be sequenced to determine the identity of the nucleotides at the novel polymorphic sites described herein. Any particular variant claimed herein could be prepared from this clone by performing *in vitro* mutagenesis using procedures well-known in the art.

UGT1A1 isogenes may be isolated using any method that allows separation of the two "copies" of the UGT1A1 gene present in an individual, which, as readily understood by the skilled artisan, may be the same allele or different alleles. Separation methods include targeted *in vivo* cloning (TIVC) in yeast as described in WO 98/01573, U.S. Patent No. 5,866,404, and U.S. Patent No. 5,972,614. Another method, which is described in U.S. Patent No. 5,972,614, uses an allele specific oligonucleotide in combination with primer extension and exonuclease degradation to generate hemizygous DNA targets. Yet other methods are single molecule dilution (SMD) as described in Ruaño et al., *Proc. Natl. Acad. Sci.* 87:6296-6300, 1990; and allele specific PCR (Ruaño et al., 1989, *supra*; Ruaño et al., 1991, *supra*; Michalatos-Beloin et al., *supra*).

The invention also provides UGT1A1 genome anthologies, which are collections of UGT1A1 isogenes found in a given population. The population may be any group of at least two individuals, including but not limited to a reference population, a population group, a family population, a clinical population, and a same sex population. A UGT1A1 genome anthology may comprise individual UGT1A1 isogenes stored in separate containers such as microtest tubes, separate wells of a microtitre plate and the like. Alternatively, two or more groups of the UGT1A1 isogenes in the anthology may be stored in separate containers. Individual isogenes or groups of isogenes in a genome anthology may be stored in any convenient and stable form, including but not limited to in buffered solutions, as DNA precipitates, freeze-dried preparations and the like. A preferred UGT1A1 genome anthology of the invention comprises a set of isogenes defined by the haplotypes shown in Table 4 below.

An isolated polynucleotide containing a polymorphic variant nucleotide sequence of the invention may be operably linked to one or more expression regulatory elements in a recombinant expression vector capable of being propagated and expressing the encoded UGT1A1 protein in a prokaryotic or a eukaryotic host cell. Examples of expression regulatory elements which may be used

include, but are not limited to, the lac system, operator and promoter regions of phage lambda, yeast promoters, and promoters derived from vaccinia virus, adenovirus, retroviruses, or SV40. Other regulatory elements include, but are not limited to, appropriate leader sequences, termination codons, polyadenylation signals, and other sequences required for the appropriate transcription and subsequent translation of the nucleic acid sequence in a given host cell. Of course, the correct combinations of expression regulatory elements will depend on the host system used. In addition, it is understood that the expression vector contains any additional elements necessary for its transfer to and subsequent replication in the host cell. Examples of such elements include, but are not limited to, origins of replication and selectable markers. Such expression vectors are commercially available or are readily constructed using methods known to those in the art (e.g., F. Ausubel et al., 1987, in "Current Protocols in Molecular Biology", John Wiley and Sons, New York, New York). Host cells which may be used to express the variant UGT1A1 sequences of the invention include, but are not limited to, eukaryotic and mammalian cells, such as animal, plant, insect and yeast cells, and prokaryotic cells, such as E. coli, or algal cells as known in the art. The recombinant expression vector may be introduced into the host cell using any method known to those in the art including, but not limited to, microinjection, electroporation, particle bombardment, transduction, and transfection using DEAE-dextran, lipofection, or calcium phosphate (see e.g., Sambrook et al. (1989) in "Molecular Cloning. A Laboratory Manual", Cold Spring Harbor Press, Plainview, New York). In a preferred aspect, eukaryotic expression vectors that function in eukaryotic cells, and preferably mammalian cells, are used. Non-limiting examples of such vectors include vaccinia virus vectors, adenovirus vectors, herpes virus vectors, and baculovirus transfer vectors. Preferred eukaryotic cell lines include COS cells, CHO cells, HeLa cells, NIH/3T3 cells, and embryonic stem cells (Thomson, J. A. et al., 1998 *Science* 282:1145-1147). Particularly preferred host cells are mammalian cells.

As will be readily recognized by the skilled artisan, expression of polymorphic variants of the UGT1A1 gene will produce UGT1A1 mRNAs varying from each other at any polymorphic site retained in the spliced and processed mRNA molecules. These mRNAs can be used for the preparation of a UGT1A1 cDNA comprising a nucleotide sequence which is a polymorphic variant of the UGT1A1 reference coding sequence shown in Figure 2. Thus, the invention also provides UGT1A1 mRNAs and corresponding cDNAs which comprise a nucleotide sequence that is identical to SEQ ID NO: 2 (Fig. 2), or its corresponding RNA sequence, except for having one or more polymorphisms selected from the group consisting of thymine at a position corresponding to nucleotide 141, guanine at a position corresponding to nucleotide 540, thymine at a position corresponding to nucleotide 1428 and thymine at a position corresponding to nucleotide 1491,and may also comprise an additional polymorphism of adenine at a position corresponding to nucleotide 211. A particularly preferred polymorphic cDNA variant comprises the coding sequence of a UGT1A1 isogene defined by haplotypes 2- 21. Fragments of these variant mRNAs and cDNAs are included in the scope of the invention, provided they contain the novel polymorphisms described herein. The invention specifically excludes polynucleotides

identical to previously identified and characterized UGT1A1 cDNAs and fragments thereof. Polynucleotides comprising a variant RNA or DNA sequence may be isolated from a biological sample using well-known molecular biological procedures or may be chemically synthesized.

As used herein, a polymorphic variant of a UGT1A1 gene fragment comprises at least one novel polymorphism identified herein and has a length of at least 10 nucleotides and may range up to the full length of the gene. Preferably, such fragments are between 100 and 3000 nucleotides in length, and more preferably between 200 and 2000 nucleotides in length, and most preferably between 500 and 1000 nucleotides in length.

In describing the UGT1A1 polymorphic sites identified herein, reference is made to the sense strand of the gene for convenience. However, as recognized by the skilled artisan, nucleic acid molecules containing the UGT1A1 gene may be complementary double stranded molecules and thus reference to a particular site on the sense strand refers as well to the corresponding site on the complementary antisense strand. Thus, reference may be made to the same polymorphic site on either strand and an oligonucleotide may be designed to hybridize specifically to either strand at a target region containing the polymorphic site. Thus, the invention also includes single-stranded polynucleotides which are complementary to the sense strand of the UGT1A1 genomic variants described herein.

Polynucleotides comprising a polymorphic gene variant or fragment may be useful for therapeutic purposes. For example, where a patient could benefit from expression, or increased expression, of a particular UGT1A1 protein isoform, an expression vector encoding the isoform may be administered to the patient. The patient may be one who lacks the UGT1A1 isogene encoding that isoform or may already have at least one copy of that isogene.

In other situations, it may be desirable to decrease or block expression of a particular UGT1A1 isogene. Expression of a UGT1A1 isogene may be turned off by transforming a targeted organ, tissue or cell population with an expression vector that expresses high levels of untranslatable mRNA for the isogene. Alternatively, oligonucleotides directed against the regulatory regions (e.g., promoter, introns, enhancers, 3′ untranslated region) of the isogene may block transcription. Oligonucleotides targeting the transcription initiation site, e.g., between positions −10 and +10 from the start site are preferred. Similarly, inhibition of transcription can be achieved using oligonucleotides that base-pair with region(s) of the isogene DNA to form triplex DNA (see e.g., Gee et al. in Huber, B.E. and B.I. Carr, Molecular and Immunologic Approaches, Futura Publishing Co., Mt. Kisco, N.Y., 1994). Antisense oligonucleotides may also be designed to block translation of UGT1A1 mRNA transcribed from a particular isogene. It is also contemplated that ribozymes may be designed that can catalyze the specific cleavage of UGT1A1 mRNA transcribed from a particular isogene.

The oligonucleotides may be delivered to a target cell or tissue by expression from a vector introduced into the cell or tissue *in vivo* or *ex vivo*. Alternatively, the oligonucleotides may be formulated as a pharmaceutical composition for administration to the patient. Oligoribonucleotides

and/or oligodeoxynucleotides intended for use as antisense oligonucleotides may be modified to increase stability and half-life. Possible modifications include, but are not limited to phosphorothioate or 2′ O-methyl linkages, and the inclusion of nontraditional bases such as inosine and queosine, as well as acetyl-, methyl-, thio-, and similarly modified forms of adenine, cytosine, guanine, thymine, and uracil which are not as easily recognized by endogenous nucleases.

Effect(s) of the polymorphisms identified herein on expression of UGT1A1 may be investigated by preparing recombinant cells and/or nonhuman recombinant organisms, preferably recombinant animals, containing a polymorphic variant of the UGT1A1 gene. As used herein, "expression" includes but is not limited to one or more of the following: transcription of the gene into precursor mRNA; splicing and other processing of the precursor mRNA to produce mature mRNA; mRNA stability; translation of the mature mRNA into UGT1A1 protein (including codon usage and tRNA availability); and glycosylation and/or other modifications of the translation product, if required for proper expression and function.

To prepare a recombinant cell of the invention, the desired UGT1A1 isogene may be introduced into the cell in a vector such that the isogene remains extrachromosomal. In such a situation, the gene will be expressed by the cell from the extrachromosomal location. In a preferred embodiment, the UGT1A1 isogene is introduced into a cell in such a way that it recombines with the endogenous UGT1A1 gene present in the cell. Such recombination requires the occurrence of a double recombination event, thereby resulting in the desired UGT1A1 gene polymorphism. Vectors for the introduction of genes both for recombination and for extrachromosomal maintenance are known in the art, and any suitable vector or vector construct may be used in the invention. Methods such as electroporation, particle bombardment, calcium phosphate co-precipitation and viral transduction for introducing DNA into cells are known in the art; therefore, the choice of method may lie with the competence and preference of the skilled practitioner. Examples of cells into which the UGT1A1 isogene may be introduced include, but are not limited to, continuous culture cells, such as COS, NIH/3T3, and primary or culture cells of the relevant tissue type, i.e., they express the UGT1A1 isogene. Such recombinant cells can be used to compare the biological activities of the different protein variants.

Recombinant nonhuman organisms, i.e., transgenic animals, expressing a variant UGT1A1 gene are prepared using standard procedures known in the art. Preferably, a construct comprising the variant gene is introduced into a nonhuman animal or an ancestor of the animal at an embryonic stage, i.e., the one-cell stage, or generally not later than about the eight-cell stage. Transgenic animals carrying the constructs of the invention can be made by several methods known to those having skill in the art. One method involves transfecting into the embryo a retrovirus constructed to contain one or more insulator elements, a gene or genes of interest, and other components known to those skilled in the art to provide a complete shuttle vector harboring the insulated gene(s) as a transgene, see e.g., U.S. Patent No. 5,610,053. Another method involves directly injecting a transgene into the embryo. A third

method involves the use of embryonic stem cells. Examples of animals into which the UGT1A1 isogenes may be introduced include, but are not limited to, mice, rats, other rodents, and nonhuman primates (see "The Introduction of Foreign Genes into Mice" and the cited references therein, In: Recombinant DNA, Eds. J.D. Watson, M. Gilman, J. Witkowski, and M. Zoller; W.H. Freeman and Company, New York, pages 254-272). Transgenic animals stably expressing a human UGT1A1 isogene and producing human UGT1A1 protein can be used as biological models for studying diseases related to abnormal UGT1A1 expression and/or activity, and for screening and assaying various candidate drugs, compounds, and treatment regimens to reduce the symptoms or effects of these diseases.

An additional embodiment of the invention relates to pharmaceutical compositions for treating disorders affected by expression or function of a novel UGT1A1 isogene described herein. The pharmaceutical composition may comprise any of the following active ingredients: a polynucleotide comprising one of these novel UGT1A1 isogenes; an antisense oligonucleotide directed against one of the novel UGT1A1 isogenes, a polynucleotide encoding such an antisense oligonucleotide, or another compound which inhibits expression of a novel UGT1A1 isogene described herein. Preferably, the composition contains the active ingredient in a therapeutically effective amount. By therapeutically effective amount is meant that one or more of the symptoms relating to disorders affected by expression or function of a novel UGT1A1 isogene is reduced and/or eliminated. The composition also comprises a pharmaceutically acceptable carrier, examples of which include, but are not limited to, saline, buffered saline, dextrose, and water. Those skilled in the art may employ a formulation most suitable for the active ingredient, whether it is a polynucleotide, oligonucleotide, protein, peptide or small molecule antagonist. The pharmaceutical composition may be administered alone or in combination with at least one other agent, such as a stabilizing compound. Administration of the pharmaceutical composition may be by any number of routes including, but not limited to oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, intradermal, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal. Further details on techniques for formulation and administration may be found in the latest edition of Remington's Pharmaceutical Sciences (Maack Publishing Co., Easton, PA).

For any composition, determination of the therapeutically effective dose of active ingredient and/or the appropriate route of administration is well within the capability of those skilled in the art. For example, the dose can be estimated initially either in cell culture assays or in animal models. The animal model may also be used to determine the appropriate concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans. The exact dosage will be determined by the practitioner, in light of factors relating to the patient requiring treatment, including but not limited to severity of the disease state, general health, age, weight and gender of the patient, diet, time and frequency of administration, other drugs being taken by the patient, and tolerance/response to the treatment.

Any or all analytical and mathematical operations involved in practicing the methods of the present invention may be implemented by a computer. In addition, the computer may execute a program that generates views (or screens) displayed on a display device and with which the user can interact to view and analyze large amounts of information relating to the UGT1A1 gene and its genomic variation, including chromosome location, gene structure, and gene family, gene expression data, polymorphism data, genetic sequence data, and clinical data population data (e.g., data on ethnogeographic origin, clinical responses, genotypes, and haplotypes for one or more populations). The UGT1A1 polymorphism data described herein may be stored as part of a relational database (e.g., an instance of an Oracle database or a set of ASCII flat files). These polymorphism data may be stored on the computer's hard drive or may, for example, be stored on a CD-ROM or on one or more other storage devices accessible by the computer. For example, the data may be stored on one or more databases in communication with the computer via a network.

Preferred embodiments of the invention are described in the following examples. Other embodiments within the scope of the claims herein will be apparent to one skilled in the art from consideration of the specification or practice of the invention as disclosed herein. It is intended that the specification, together with the examples, be considered exemplary only, with the scope and spirit of the invention being indicated by the claims which follow the examples.

EXAMPLES

The Examples herein are meant to exemplify the various aspects of carrying out the invention and are not intended to limit the scope of the invention in any way. The Examples do not include detailed descriptions for conventional methods employed, such as in the performance of genomic DNA isolation, PCR and sequencing procedures. Such methods are well-known to those skilled in the art and are described in numerous publications, for example, Sambrook, Fritsch, and Maniatis, "Molecular Cloning: A Laboratory Manual", 2nd Edition, Cold Spring Harbor Laboratory Press, USA, (1989).

EXAMPLE 1

This example illustrates examination of various regions of the UGT1A1 gene for polymorphic sites.

Amplification of Target Regions

The following target regions of the CSF1R gene were amplified using PCR primer pairs. The primers used for each region are represented below by providing the nucleotide positions of their initial and final nucleotides, which correspond to positions in Figure 1.

Primer Pairs

Fragment No.	Forward Primer	Reverse Primer	PCR Product
Fragment 1	2301-2324	complement of 2992-2969	692 nt
Fragment 2	2591-2611	complement of 3173-3152	583 nt
Fragment 3	2837-2859	complement of 3465-3444	629 nt
Fragment 4	3152-3173	complement of 3717-3693	566 nt
Fragment 5	9019-9038	complement of 9720-9700	702 nt
Fragment 6	9863-9884	complement of 10549-10530	687 nt
Fragment 7	10365-10386	complement of 11029-11009	665 nt
Fragment 8	14388-14410	complement of 14875-14852	488 nt
Fragment 9	14515-14537	complement of 15117-15094	603 nt

PCR Primer Pairs

These primer pairs were used in PCR reactions containing genomic DNA isolated from immortalized cell lines for each member of the Index Repository. The PCR reactions were carried out under the following conditions:

Reaction volume	= 10 µl
10 x Advantage 2 Polymerase reaction buffer (Clontech)	= 1 µl
100 ng of human genomic DNA	= 1 µl
10 mM dNTP	= 0.4 µl
Advantage 2 Polymerase enzyme mix (Clontech)	= 0.2 µl
Forward Primer (10 µM)	= 0.4 µl
Reverse Primer (10 µM)	= 0.4 µl
Water	= 6.6µl

Amplification profile:
97°C - 2 min. 1 cycle

97°C - 15 sec.
70°C - 45 sec. } 10 cycles
72°C - 45 sec.

97°C - 15 sec.
64°C - 45 sec. } 35 cycles
72°C - 45 sec.

Sequencing of PCR Products

The PCR products were purified using a Whatman/Polyfiltronics 100 µl 384 well unifilter plate essentially according to the manufacturers protocol. The purified DNA was eluted in 50 µl of distilled water. Sequencing reactions were set up using Applied Biosystems Big Dye Terminator chemistry essentially according to the manufacturers protocol. The purified PCR products were sequenced in both directions using the primer sets described previously or those represented below by the nucleotide positions of their initial and final nucleotides, which correspond to positions in Figure 1. Reaction products were purified by isopropanol precipitation, and run on an Applied Biosystems 3700 DNA Analyzer.

Sequencing Primer Pairs

Fragment No.	Forward Primer	Reverse Primer
Fragment 1	2478-2499	complement of 2947-2928
Fragment 2	2685-2704	complement of 3143-3124
Fragment 3	2932-2951	complement of 3380-3361
Fragment 4	3173-3192	complement of 3655-3634
Fragment 5	9183-9204	complement of 9623-9605
Fragment 6	10016-10035	complement of 10450-10431
Fragment 7	10427-10448	complement of 10964-10944
Fragment 8	14416-14434	complement of 14843-14824
Fragment 9	14600-14619	complement of 15027-15007

Analysis of Sequences for Polymorphic Sites

Sequences were analyzed for the presence of polymorphisms using the Polyphred program (Nickerson et al., *Nucleic Acids Res.* 14:2745-2751, 1997). The presence of a polymorphism was confirmed on both strands. The polymorphisms and their locations in the UGT1A1 gene are listed in Table 2 below.

Table 2. Polymorphic Sites Identified in the UGT1A1 Gene

Polymorphic Site Number	PolyId[a]	Nucleotide Position in GenBank	Nucleotide Position	Reference Allele	Variant Allele
PS1	98916	66261(Acc#AC006985.2)	2510	T	C
PS2	14143	66015(Acc#AC006985.2)	2756	C	T
PS3[R]	14144	65945(Acc#AC006985.2)	2826	G	A
PS4	14142	65616(Acc#AC006985.2)	3155	A	G
PS5	14111	65203(Acc#AC006985.2)	3568	C	T
PS6	98908	59263(Acc#AC006985.2)	9508	T	C
PS7	98907	59260(Acc#AC006985.2)	9511	C	T
PS8	14136	58680(Acc#AC006985.2)	10091	C	T
PS9	14137	58677(Acc#AC006985.2)	10094	C	T
PS10	14138	58676(Acc#AC006985.2)	10095	T	C
PS11	14140	58631(Acc#AC006985.2)	10140	T	C
PS12	14148	54348(Acc#AC006985.2)	14423	A	T
PS13	14147	54058(Acc#AC006985.2)	14713	C	T
PS14	14146	53995(Acc#AC006985.2)	14776	C	T
PS15	14145	53800(Acc#AC006985.2)	14971	T	C

[a]PolyId is a unique identifier assigned to each PS by Genaissance Pharmaceuticals, Inc.
[R] Previously reported in the literature

EXAMPLE 2

This example illustrates analysis of the UGT1A1 polymorphisms identified in the Index Repository for human genotypes and haplotypes.

The different genotypes containing these polymorphisms that were observed in the reference population are shown in Table 3 below, with the haplotype pair indicating the combination of haplotypes determined for the individual using the haplotype derivation protocol described below. In

Table 3, homozygous positions are indicated by one nucleotide and heterozygous positions are indicated by two nucleotides. Missing nucleotides in any given genotype in Table 3 were inferred based on linkage disequilibrium and/or Mendelian inheritance.

Table 3. Genotypes and Haplotype Pairs Observed for UGT1A1 Gene

Genotype	Polymorphic Sites															
Number	PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	PS9	PS10	PS11	PS12	PS13	PS14	PS15	HAP PAIR
1	T	C	G	A	C	T	C	C	C	T	T	A	C	C	T	1 1
2	T	C	G	A	C	T	C	C	C	T/C	T	A	C	C	T	1 2
3	T	C	G/A	A	C	T	C	C	C	T	T	A	C	C	T	1 3
4	T	C	G	A	C/T	-	-	C	C	T/C	T	A	C	C	T	1 4
5	T	C	G	A	C	T	C	C	C/T	T/C	T	A	C	C	T	1 6
6	-	C	G	A	C	T	C	C	C	T/C	T	A/T	C	C	T	1 7
7	T/C	C	G	A	C	T	C/T	C	C/T	T/C	T	A/T	C	C	T	1 8
8	T	C	G	A	C	T	C	C/T	C	T	T/C	A	C	C	T	1 9
9	T	C	G	A	C	T	C	C	C/T	T/C	T	-	C/T	C	T	1 10
10	T	C/T	G	A	C	T	C	C	C	T/C	T	A	C	C	T	1 13
11	T	C	G	A	C	T	C	C	C	T/C	T/C	A	C	C	T	1 14
12	T	C	G	A	C	T/C	C	C	C	T/C	T	A	C	C	T	1 15
13	T	C	G	A	C	T/C	C	C	C	T	T	A	C	C	T	1 16
14	T	C	G	A	C	T	C	C	C	T	T/C	A	C	C	T	1 17
15	T	C	G	A	C	T	C	C	C	T	T	A	C	C	T/C	1 18
16	T	C	G	A/G	C	T	C	C	C	T	T	A	C	C	T	1 19
17	T	C	G	A	C	T	C	C	C	T	T	A	C	C/T	T	1 20
18	T	C	G	A	C	T	C	C	C	C	T	A	C	C	T	2 2
19	T	C	G	A	C	T	C	C	C	C	T	-	C	C	-	2 2
20	T	C	G	A	C	T	C	C	C/T	C	T	A	C	C	T	2 6
21	T/C	C	G	A	C	T	C/T	C	C	C	T	A	C/T	C	T	2 11
22	T	C	G	A	C	T	C/T	C	C/T	C	T	A	C	C	T	2 12
23	T/C	C	G	A	C	T	C/T	C	C/T	C	T	A	C	C	T	2 21
24	C	C	G	A	C	T	T	C	C	C	T	A	C	C	T	5 5

The haplotype pairs shown in Table 3 were estimated from the unphased genotypes using a computer-implemented extension of Clark's algorithm (Clark, A.G. 1990 *Mol Bio Evol* 7, 111-122) for assigning haplotypes to unrelated individuals in a population sample. In this method, haplotypes are assigned directly from individuals who are homozygous at all sites or heterozygous at no more than one of the variable sites. This list of haplotypes is augmented with haplotypes obtained from two families (one three-generation Caucasian family and one two-generation African-American family) and then used to deconvolute the unphased genotypes in the remaining (multiply heterozygous) individuals.

By following this protocol, it was determined that the Index Repository examined herein and, by extension, the general population contains the 21 human UGT1A1 haplotypes shown in Table 4 below.

Table 4. Haplotypes Identified in the UGT1A1 Gene

Haplotype	Polymorphic Sites														
Number	PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	PS9	PS10	PS11	PS12	PS13	PS14	PS15
1	T	C	G	A	C	T	C	C	C	T	T	A	C	C	T
2	T	C	G	A	C	T	C	C	C	C	T	A	C	C	T
3	T	C	A	A	C	T	C	C	C	T	T	A	C	C	T
4	T	C	G	A	T	T	C	C	C	C	T	A	C	C	T
5	C	C	G	A	C	T	T	C	C	C	T	A	C	C	T
6	T	C	G	A	C	T	C	C	T	C	T	A	C	C	T
7	T	C	G	A	C	T	C	C	C	C	T	T	C	C	T
8	C	C	G	A	C	T	T	C	T	C	T	T	C	C	T
9	T	C	G	A	C	T	C	T	C	T	C	A	C	C	T
10	T	C	G	A	C	T	C	C	T	C	T	A	T	C	T
11	C	C	G	A	C	T	T	C	C	C	T	A	T	C	T
12	T	C	G	A	C	T	T	C	T	C	T	A	C	C	T
13	T	T	G	A	C	T	C	C	C	C	T	A	C	C	T
14	T	C	G	A	C	T	C	C	C	C	C	A	C	C	T
15	T	C	G	A	C	C	C	C	C	C	T	A	C	C	T
16	T	C	G	A	C	C	C	C	C	T	T	A	C	C	T
17	T	C	G	A	C	T	C	C	C	T	C	A	C	C	T
18	T	C	G	A	C	T	C	C	C	T	T	A	C	C	C
19	T	C	G	G	C	T	C	C	C	T	T	A	C	C	T
20	T	C	G	A	C	T	C	C	C	T	T	A	C	T	T
21	C	C	G	A	C	T	T	C	T	C	T	A	C	C	T

The number of chromosomes in unrelated individuals characterized by a given haplotype within each ethnic group in the reference population is indicated in Table 5 below, using the following abbreviations: AF (African descent), AS (Asian), CA (Caucasian), HL (Hispanic/Latino), and AM (Native American).

Table 5: Frequencies of Observed HAPs in the UGT1A1 Gene						
HAP No	AF	AS	CA	HL	AM	Total
1	20	31	36	30	6	123
2	11	0	0	0	0	11
3	0	5	0	1	0	6
4	0	0	4	1	0	5
5	2	0	0	0	0	2
6	2	0	0	0	0	2
7	0	0	0	1	0	1
8	1	0	0	0	0	1
9	0	0	0	1	0	1
10	1	0	0	0	0	1
11	1	0	0	0	0	1
12	1	0	0	0	0	1
13	0	0	1	0	0	1
14	0	0	0	1	0	1
15	0	1	0	0	0	1
16	0	1	0	0	0	1
17	0	0	1	0	0	1
18	0	1	0	0	0	1
19	0	0	0	1	0	1
20	0	1	0	0	0	1
21	1	0	0	0	0	1

The number of unrelated individuals having a given haplotype pair within each ethnic group in the reference population is indicated in Table 6 below, using the aforementioned abbreviations as in Table 5.

Table 6: Frequencies of Observed HAP Pairs in the UGT1A1 Gene							
HAP	PAIR	AF	AS	CA	HL	AM	Total
1	1	6	11	15	12	3	47
2	1	5	0	0	0	0	5
2	2	1	0	0	0	0	1
3	1	0	5	0	1	0	6
4	1	0	0	4	1	0	5
5	5	1	0	0	0	0	1
6	1	1	0	0	0	0	1
6	2	1	0	0	0	0	1
7	1	0	0	0	1	0	1
8	1	1	0	0	0	0	1
9	1	0	0	0	1	0	1
10	1	1	0	0	0	0	1
11	2	1	0	0	0	0	1
12	2	1	0	0	0	0	1
13	1	0	0	1	0	0	1
14	1	0	0	0	1	0	1
15	1	0	1	0	0	0	1
16	1	0	1	0	0	0	1
17	1	0	0	1	0	0	1
18	1	0	1	0	0	0	1
19	1	0	0	0	1	0	1
20	1	0	1	0	0	0	1
21	2	1	0	0	0	0	1

In view of the above, it will be seen that the several advantages of the invention are achieved and other advantageous results attained.

As various changes could be made in the above methods and compositions without departing from the scope of the invention, it is intended that all matter contained in the above description and shown in the accompanying drawings shall be interpreted as illustrative and not in a limiting sense.

All references cited in this specification, including patents and patent applications, are hereby incorporated in their entirety by reference. The discussion of references herein is intended merely to summarize the assertions made by their authors and no admission is made that any reference constitutes prior art. Applicants reserve the right to challenge the accuracy and pertinency of the cited references.

What is Claimed is:

1. A method for haplotyping the UDP glycosyltransferase 1 (UGT1A1) gene of an individual which comprises determining whether the individual has one of the UGT1A1 haplotypes shown in Table 4 or one of the haplotype pairs shown in Table 3.

2. The method of claim 1, wherein the determining step comprises identifying the phased sequence of nucleotides present at each of PS1-15 on at least one copy of the individual's UGT1A1 gene.

3. The method of claim 1, wherein the determining step comprises identifying the phased sequence of nucleotides present at each of PS1-15 on both copies of the individual's UGT1A1 gene.

4. A method for genotyping the UDP glycosyltransferase 1 (UGT1A1) gene of an individual, comprising determining for the two copies of the UGT1A1 gene present in the individual the identity of the nucleotide pair at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15.

5. The method of claim 4, wherein the determining step comprises:
   (a) isolating from the individual a nucleic acid mixture comprising both copies of the UGT1A1 gene, or a fragment thereof, that are present in the individual;
   (b) amplifying from the nucleic acid mixture a target region containing the selected polymorphic site;
   (c) hybridizing a primer extension oligonucleotide to one allele of the amplified target region;
   (d) performing a nucleic acid template-dependent, primer extension reaction on the hybridized genotyping oligonucleotide in the presence of at least two different terminators of the reaction, wherein said terminators are complementary to the alternative nucleotides present at the selected polymorphic site; and
   (e) detecting the presence and identity of the terminator in the extended genotyping oligonucleotide.

6. The method of claim 4, which comprises determining for the two copies of the UGT1A1 gene present in the individual the identity of the nucleotide pair at each of PS1-15.

7. A method for haplotyping the UDP glycosyltransferase 1 (UGT1A1) gene of an individual which comprises determining, for one copy of the UGT1A1 gene present in the individual, the identity of the nucleotide at two or more polymorphic sites selected from the group consisting of PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15.

8. The method of claim 7, further comprising determining the identity of the nucleotide at PS3

9. The method of claim 7, wherein the determining step comprises:
   (a) isolating from the individual a nucleic acid sample containing only one of the two copies of the UGT1A1 gene, or a fragment thereof, that is present in the individual;

(b) amplifying from the nucleic acid molecule a target region containing the selected polymorphic site;

(c) hybridizing a primer extension oligonucleotide to one allele of the amplified target region;

(d) performing a nucleic acid template-dependent, primer extension reaction on the hybridized genotyping oligonucleotide in the presence of at least two different terminators of the reaction, wherein said terminators are complementary to the alternative nucleotides present at the selected polymorphic site; and

(e) detecting the presence and identity of the terminator in the extended genotyping oligonucleotide.

10. A method for predicting a haplotype pair for the UDP glycosyltransferase 1 (UGT1A1) gene of an individual comprising:

(a) identifying a UGT1A1 genotype for the individual, wherein the genotype comprises the nucleotide pair at two or more polymorphic sites selected from the group consisting of PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15;

(b) enumerating all possible haplotype pairs which are consistent with the genotype;

(c) comparing the possible haplotype pairs to the data in Table 3; and

(d) assigning a haplotype pair to the individual that is consistent with the data.

11. The method of claim 10, wherein the identified genotype of the individual comprises the nucleotide pair at each of PS1-15.

12. A method for identifying an association between a trait and at least one haplotype or haplotype pair of the UDP glycosyltransferase 1 (UGT1A1) gene which comprises comparing the frequency of the haplotype or haplotype pair in a population exhibiting the trait with the frequency of the haplotype or haplotype pair in a reference population, wherein the haplotype is selected from haplotypes 1-21 shown in Table 4 and the haplotype pair is selected from the haplotype pairs shown in Table 3, wherein a higher frequency of the haplotype or haplotype pair in the trait population than in the reference population indicates the trait is associated with the haplotype or haplotype pair.

13. The method of claim 12, wherein the trait is a clinical response to a drug targeting UGT1A1.

14. A composition comprising at least one genotyping oligonucleotide for detecting a polymorphism in the UDP glycosyltransferase 1 (UGT1A1) gene at a polymorphic site selected from the group consisting of PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15.

15. The composition of claim 14, wherein the genotyping oligonucleotide is an allele-specific oligonucleotide that specifically hybridizes to an allele of the UGT1A1 gene at a region containing the polymorphic site.

16. The composition of claim 15, wherein the allele-specific oligonucleotide comprises a nucleotide sequence selected from the group consisting of SEQ ID NOS:4-17, the complements of SEQ ID

NOS:4-17, and SEQ ID NOS:18-45.

17. The composition of claim 14, wherein the genotyping oligonucleotide is a primer-extension oligonucleotide.

18. The composition of claim 17, wherein the primer extension oligonucleotide comprises a nucleotide sequence selected from the group consisting of SEQ ID NOS:46-73.

19. A kit for genotyping the UGT1A1 gene of an individual, which comprises a set of oligonucleotides designed to genotype each of PS1, PS2, PS4, PS5, PS6, PS7, PS8, PS9, PS10, PS11, PS12, PS13, PS14 and PS15.

20. The kit of claim 19, which further comprises oligonucleotides designed to genotype PS3.

21. An isolated polynucleotide comprising a nucleotide sequence selected from the group consisting of:

(a) a first nucleotide sequence which is a polymorphic variant of a reference sequence for the UDP glycosyltransferase 1 (UGT1A1) gene or a fragment thereof, wherein the reference sequence comprises SEQ ID NO:1 and the polymorphic variant comprises a UGT1A1 isogene defined by a haplotype selected from the group consisting of haplotypes 1-21 in Table 4; and

(b) a second nucleotide sequence which is complementary to the first nucleotide sequence.

22. The isolated polynucleotide of claim 21, which is a DNA molecule and comprises both the first and second nucleotide sequences and further comprises expression regulatory elements operably linked to the first nucleotide sequence.

23. A recombinant nonhuman organism transformed or transfected with the isolated polynucleotide of claim 21, wherein the organism expresses a UGT1A1 protein encoded by the first nucleotide sequence.

24. The recombinant organism of claim 23, which is a nonhuman transgenic animal.

25. The isolated polynucleotide of claim 21, wherein the first nucleotide sequence is a polymorphic variant of a fragment of the UGT1A1 gene, the fragment comprising one or more polymorphisms selected from the group consisting of cytosine at PS1, thymine at PS2, guanine at PS4, thymine at PS5, cytosine at PS6, thymine at PS7, thymine at PS8, thymine at PS9, cytosine at PS10, cytosine at PS11, thymine at PS12, thymine at PS13, thymine at PS14 and cytosine at PS15.

26. An isolated polynucleotide comprising a nucleotide sequence which is a polymorphic variant of a reference sequence for the UGT1A1 cDNA or a fragment thereof, wherein the reference sequence comprises SEQ ID NO:2 and the polymorphic variant comprises the coding sequence of a UGT1A1 isogene defined by one of the haplotypes shown in Table 4.

27. A recombinant nonhuman organism transformed or transfected with the isolated polynucleotide of claim 26, wherein the organism expresses a UDP glycosyltransferase 1 (UGT1A1) protein encoded by the polymorphic variant sequence.

28. The recombinant organism of claim 27, which is a nonhuman transgenic animal.

29. A computer system for storing and analyzing polymorphism data for the UDP glycosyltransferase 1 gene, comprising:

    (a) a central processing unit (CPU);

    (b) a communication interface;

    (c) a display device;

    (d) an input device; and

    (e) a database containing the polymorphism data;

    wherein the polymorphism data comprises the genotypes and haplotype pairs shown in Table 3 and the haplotypes shown in Table 4.

30. A genome anthology for the UDP glycosyltransferase 1 (UGT1A1) gene which comprises UGT1A1 isogenes defined by any one of haplotypes 1-21 shown in Table 4.

POLYMORPHISMS IN THE UGT1A1 GENE

```
GAATTCAAGG GATTCAAGGA AGGTGGCTTT GCTTCCCGGG AGGGTCCTGT
AGATGATCTA CAGGGCACTG GACATGTTTA TGTTGCTCCT TTAGTAATAA    100
GCCTGTCATT CTGATTTGAT GAAAGGAGAT GAAAGGAGCT GGTAGTGTGT
CTGATGGTGG CCTACTAACT TATGTCTTCA GCTTAAAAAG AAAGTAGCTT    200
CAAAAGGGTT CCAGAAACAC TTTCCATGGA CGTGTCACTC TTTAGCAGCC
CCCAAAGCAA GACCATCATA TTGCTGCCCT GCTGTGTGAT TTCTCAGCCC    300
CTAGAGCACC ATCCCCTGTA ATTGCCTGGT CATGAGTTTG TCTCTGTCTA
CCTGACCCCT CCTTTCAGGC AAGGACCATT TCTAACTTGA CTTTCTGGGC    400
CTAGTTCCTA GCATAGTGAC TGCCATCCAG TAGGGCTCAC ACGTTCCATA
AATATTTGGC AGATGAGGGA ATTAGCAATG GGTTCTGCTT TGGTTTCAGA    500
GCAGATATTA ATTGGATTGC TTAGTAGTGG TTCTCTGTTG TAATTCATGA
GCATGAATGT GGATTGCCCA CTATTCAGAT TAGTAAGTAT TTCTTGGTCA    600
AGGGCAGAGC TGTGGCCACA AACCATCCAG GTACACAGCA GAAGCAGCCT
CAAAAAGCTT GGAAGCTCTG CATGATGCAG GAAAGTCATA AAATCATTAC    700
AGTGGTGACT TATGTGTTTA TAGCCCCTTT ACTGTCTATA ATCTGCAAAT
GAACTCACAC AGCATTGGGA CTTTGGAAGA ATTATCACCC TTAAGGTTTA    800
AATTAAACTG TGAATTTCAG AATTTCTAAT AAGGACACAA CAAAGAGTGA
AAGCATTGCT ATGTCTATTC TGCTTGCCCA GAATCTTGGT CCTAAAAAAT    900
GAAGAGTGTT TGGGTGTGGG GAGGAGCTTC AGTGTGCATG TGCATGCAAA
GTACCTACTC TAAGGAGAAG AATGAGAGGG TACCCTAATT ACCTGTTAAT    1000
ATGTCCCATA GGACACCAAA ACTCTAGTTA GCTGTTTCTC TATGATCCTC
TAAGCACATC CCCAAGTATG GCTGGCCAGT GATGTGTATG GTTCAAATGT    1100
TGGGATCTGT GCAGTTATCT TGGAATTGTA TAGTACAGCA GTATATCCCC
CCCAAAAAGA GTGTAATACT TCCAATTCTG GCTGCACAAT ACTTGCCCCA    1200
TAGTCCATGG TCAATAAATA CAAATTTGAG TTGTTTTTGC TCATCTTTCC
CTTTTGACTT CAAATCAGTC ATCAGAATTT CCCCAAATGC CTTTCCCCTG    1300
GATCTTGGGC CAGTGGAATG AGTACAATTT AACTTAATTG AATTTGCTTA
TCTATTTGGT TTCCTGTTGT GAACAAAAGT TCTCTGAAAA GGAATTTGGA    1400
AGAAAGAGAC TTTGTTCTAG TGAACAGTTT GCAAACCAGG GAGTTACAGC
CTCTGGTACG CAATGAAGGT GAGTTCCACA GAACACAAGG CAGGCAGGTT    1500
TCACGGCAAA AAGTTCCTTC CCAGGTTCCC AATCAGGTCC ATTTATGCAA
ATGAAGGATG GAAACTTGCT TAGTTCTTAT TGGTCACTGC AGCTGCATTC    1600
TGATTGGTTG ATGAAGCTGA GCCCTGAGTG GCTGAGGTGG GTGAGCTTTA
ATTGGTTGGT TCAGGTGAGC GCTGAAAATC TCAACTATAA AAAGGTACAG    1700
GTTTTCAGGA TACTCAGAGT AACCGTGTGA CCTGTAGTAA GCAAAGGGCC
AGTTGGCTCT ATTTTAAATC CAGGCCCAGT TAGCCACTCA AGATCTATCT    1800
TACAGGACTG GCTCTTTCAG GTTCACACTA ATAAAGGCCT GTCCTTGGGG
AAGACTTCTG TTCACATGCG CTCCAGTGAA TTTCCCTTTC TGGTCATTCT    1900
CTACCCCAGC ACGCCCCCCA CCCCCGACCC GCCCCACCCA CCCACCTGTT
CATTTCCTTC TTAGCATGCT TCACGATTTC TAAGTTCCTG CTCATGTGTT    2000
TAAATTGTGA GTCTGGCTCA CCTCATGGCG CGTGCTCGTG TGGTGGGCTC
TGCTGCAGCC TCAAGACCCC ACACTGTGCT GGACTCAATA AATATTGTTG    2100
GACGAAGGAA TGAAACACAT GATACAAGTG AGCAGGCAGT ACCGGGGGAG
CTGTGGAGTG GGCACTCTTA CAGGTTTCCA TGGCGAAAGC GGGGGTACAG    2200
TTGTGTTCTT TTCTTTCTAA AAGGCTTTCT AAAAAGCCTT CTGTTTAATT
TCTGGAAAAG AAGCCTAACT TGTTCACTAC ATAGTCGTCC TTCTTCCTCT    2300
CTGGTAACAC TTGTTGGTCT GTGGAAATAC TAATTTAATG GATCCTGAGG
TTCTGGAAGT ACTTTGCTGT GTTCACTCAA GAATGTGATT TGAGTATGAA    2400
ATTCCAGCCA GTTCAACTGT TGTTGCCTAT TAAGAAACCT AATAAAGCTC
CACCTTCTTT ATCTCTGAAA GTGAACTCCC TGCTACCTTT GTGGACTGAC    2500
AGCTTTTTAT AGTCACGTGA CACAGTCAAA CATTAACTTG GTGTATCGAT
         C
```

FIGURE 1A

```
TGGTTTTTGC CATATATATA TATATAAGTA GGAGAGGGCG AACCTCTGGC    2600
AGGAGCAAAG GCGCCATGGC TGTGGAGTCC CAGGGCGGAC GCCCACTTGT
    [EXON 1: 2616..
CCTGGGCCTG CTGCTGTGTG TGCTGGGCCC AGTGGTGTCC CATGCTGGGA    2700
AGATACTGTT GATCCCAGTG GATGGCAGCC ACTGGCTGAG CATGCTTGGG
GCCATCCAGC AGCTGCAGCA GAGGGGACAT GAAATAGTTG TCCTAGCACC    2800
    T
TGACGCCTCG TTGTACATCA GAGACGGAGC ATTTTACACC TTGAAGACGT
                        A
ACCCTGTGCC ATTCCAAAGG GAGGATGTGA AAGAGTCTTT TGTTAGTCTC    2900
GGGCATAATG TTTTTGAGAA TGATTCTTTC CTGCAGCGTG TGATCAAAAC
ATACAAGAAA ATAAAAAAGG ACTCTGCTAT GCTTTTGTCT GGCTGTTCCC    3000
ACTTACTGCA CAACAAGGAG CTCATGGCCT CCCTGGCAGA AAGCAGCTTT
GATGTCATGC TGACGGACCC TTTCCTTCCT TGCAGCCCCA TCGTGGCCCA    3100
GTACCTGTCT CTGCCCACTG TATTCTTCTT GCATGCACTG CCATGCAGCC
TGGAATTTGA GGCTACCCAG TGCCCCAACC CATTCTCCTA CGTGCCCAGG    3200
   G
CCTCTCTCCT CTCATTCAGA TCACATGACC TTCCTGCAGC GGGTGAAGAA
CATGCTCATT GCCTTTTCAC AGAACTTTCT GTGCGACGTG GTTTATTCCC    3300
CGTATGCAAC CCTTGCCTCA GAATTCCTTC AGAGAGAGGT GACTGTCCAG
GACCTATTGA GCTCTGCATC TGTCTGGCTG TTTAGAAGTG ACTTTGTGAA    3400
GGATTACCCT AGGCCCATCA TGCCCAATAT GGTTTTTGTT GGTGGAATCA
ACTGCCTTCA CCAAAATCCA CTATCCCAGG TGTGTATTGG AGTGGGACTT    3500
                ..3479]
TTACATGCGT ATATTCTTTC AGATGTATTA CTTTGGATCG ATTAACTAGC
CCCAGATATA TGCTGAGCAA GCATTCTGAG ATAATTTAAA ATGCCCTCTT    3600
                T
TTGTTAATTT TTGACTCCTA GGTTTGAGTC TGTCTTTGGC ATCATCTTCT
GGATGATTTC TTGGTATCTG AGATTTCGGG AAAGCATTCC TTGGACATTT    3700
TACTCTGTGT GCTCCAGTGG ATAGTAATCA ATTAGAAACA ACAAGCTGTT
AAATGCCATA GGCACAGAAT GCTGGGTTTG GGGCACCCTG CAGAAAACTC    3800
AGTTGAAGCC TGCACCTTGC CCTGGATTCA GTCAGGCAGG CAATGTTCAG
GACTGATGAA ATCATTCTTT GATGATGATA GATCCTGGAA ATGAAAGTTG    3900
CCTTTGTGAC CCTGGTTAAA GCTCCAGTTT CTAAATATTC TGATAAGAAG
CTAAATCCTG CAGTCCGTTC TCTTCTAATG AGTGAATCAC CAGACAGTCA    4000
GGTTCTGACA TGATACAGAA AGGTTGTAGG TTTCATTCTC AAGCTATTAG
GTTTATTTTT CCCCTACAGA GTTTGAAGTA TGCAAAAAGT AGCATTCACA    4100
TCCTCATCGA AATCTCAGCA GAGGATAGAA AAGAACAGGA GAGGCTCCTT
CAGATGGAGC GTTAGGGAAT TACTCTTTGA GGAGGTGACA TTTCAGAGAG    4200
CGTTCATTCA CTTATCCTGC AAAGATTGGC TGAGGATCTA CTGGCAGCCC
AGGCACTTCC CAGGTGCTGC GTCTGGCTCC CATTAAGGGG ACTGATATCA    4300
CCTTCGGAGG TGACCTTATT TCCACTATAC CTCCAATGTG ATTTGTATTT
TATTTTTTTT AATTTTCTGT GCATTTTCCT TCATAGCACA TCAAATATGG    4400
CAGCCATTTC ACTTAGATAG TTGTTGATTG TCCGCTTCAC ATCATGAGCC
ATGTGGGGAC CTGTGTGACT TTGCATTAAT CACATCCACT GTATGCGGCG    4500
TCCTCAACAC CTGCCAATGG GTCTGCATGT ATTTGGCGCC CCATAAATCT
CAGCACCTAA GGCACAGAAT AGGCACCCAC CGAATATGTG TTACATTAAT    4600
GAATGAGAAG AAAGGTGCCA ACCGAGGTCT AGTTAATGGG TCGAGAGTAA
TCCACAATAG CTCTTTTTAG TTCTTTGTAC TCCAGCTATT ACATACCAAT    4700
ATGTATATAG AAACATATGT AAAATTTTTT GGTTGCTTTT TCTACAAAAT
AGAGTAACAG TGTATTCCCA CTGCCCACTT ACCGATAATG TCATGGATAT    4800
CACTCCAGTT TTAAATGCTA TTACTTTTTA AACTATGAAA TAGTATTTCA
TGGTACTTGT GTACCACAGT GTATTCTGCT GGAGATCTAG TCTAGTTCCC    4900
CACAGAGGAA CATTACAATT TGTATTCCAG GAGTTTTGTT GTTGTGACCT
```

FIGURE 1B

```
CAAACACTTC CTTTAAAAAG ATAAGCTATT TTGTAGTTTA AAAAACATTT 5000
GTTCTGTTTC TTTCTCATTC ATCTTTTCTT AAGTATTTTA CACGGTTTTT
TTTTTTTGGT CACTACTGTG AATGTGTTAT TTTTTTGCAT TTCTATCTCT 5100
AGCTGATTAT CTACTCATTA CTCAGCTATC TCATCAAAAT ATTGATTTTC
ATAATAAAAA ATAATAGGCA GTCATTTGCT GATAAAGAAA TTTTGGTTTC 5200
TTCTCTTATA AATTCCATGC CAAATATCAG GGCTATTGAA TTTATTAGAA
TCTCTAAAAA CAGTTGAATA ATTCTGGCAA TAGGAAAGAT GCCCGTCTTG 5300
CTGCTATTTT AGTGGAAATT GATTATCATT TCATTATTTT GCATTATGTT
AGCCATTGTT TTCTGAACAG GCTTTATTGA TTTAGATAAT TTCCTTCTTT 5400
GCGTGAGGAT GTTTGTAGGA GAGGCACCGA ACTTTATCAG CTGCCTTTCT
GGCATTTATT GATATAACCA TAAAAGTCTA AGTGGTGAAC TGTGTTGACT 5500
ACATATTTGT TGTTGCCTTG TTTGGTGCAG TCAGGCTTAG GTGTGAAAAT
ATGTTTTTAA ATTGTACCTT TTAGTAACCT GTTTTGTCTT GTTGCATGTT 5600
TTAATCTGAA ATTCCACTTT TTGGATATTA ATATTACCAC TTCTGTATTA
TTTTTGTTTA CATTTCCCTA GCACATCTTT AGTACTCCTT TGTCTTCAAG 5700
CTTTCTTCCT TTTTAAACAA CATGGCACTG GTATTTTTAA TCCAGTCAGG
CAGTTGCTTT AATAAGTGCA TTTTGCCTAT TTGAATCTAA CAATTAATAG 5800
ATTTGATTGT AACTCTCTCA GTTTACTTTA TGTTTAGTTG ACTTTGCCAT
TCTCCTTTTT CCGGATTTCT ACTGGTTGGT CAAGTTACTG TTCTTATTTT 5900
CTCTTTCTTC CTTTGTTAAC TAAAAATGCC ACTCTGCACT ACCATTCCTC
TTGTGTTGAT GGTCCTATTC TCAATACTCT TGATAAAACT CCTGAACTTT 6000
AAGAATAAAG ATAAAACTTT TATTGCACAA AGAAGTCCAT AGAGAAAGCA
CAACCTGGCA TTGGCGTGTC TTTGGTGTGT CTGAAGGAAA AGAGATAGTG 6100
GAACAACATT GGGAGAAAAG GAATGAAACT CAAGAATTCC AAGATGTTCC
TCCCCTGCCA GGGTAAGATA GCAGTGGTTC ACAGACAATC GCAATGCTGG 6200
GTCTGAGAAA AATAACTAAA CAGAAGATTA GTGAGGACCA AGGCTTCGAG
ATGGCCAGGA GAGGAAAGCT TGGGAGCAGG GAAGGTTGAG ATATATGTGG 6300
GTTACTGGGA ATGCGTGATG GTGAAGTCAC AGATGACCCA CATGGTGTCT
AAGTGCTAAA GAAGAATTCT GGGAAAATGA AATGCATTTG GGAAGGGAAA 6400
ATCTAATTAA AAGCCTAAAC TAAAAATACA AAATTCTTGG TAAAGTTTAG
GAGTTATGTT AAATGTCTCA TTTTGGCTGG TGAAGTCTCA TCAGAACAGG 6500
GAAATTCTCT CATTCAGGGG CATCTCATCT TTTCTTTGAA GGGAATCAAT
GGTGGGGGAT TGGAGTGTTA TTTTCAGTTA ATATGTTGCT TCACTCTTTG 6600
GTCATTCCGG TAACTGTGAA GTCAGGGTGA AGTTTAAGGG AAGCTTTGCC
AAGTAGGGGA TGGACTTCAC CTTTATTGAG CCTCATAGTA GCTGGCTCAG 6700
GTAGGAGTTG GCCGTGATGA CAACTTCTCT GCAGTTTGCC CTGCGTGAAT
CTCCAGATGA ACTTTTGTGC CATTTAAACT TTCGTGATCT CCTGCTATTT 6800
AACTTCGAAT GTTTATGGAC CTGTGGGTTC AATTTTGTGT GAATCACATC
CTGCTGATTG CTGAGTGGGC GTGTGGGAGG GTGTGCCTGG AGGAGAACTT 6900
AGACTCGGCC TTTTCCAGAT GAGCTTCAGT GTAAGAGTGG GTTTCATGAA
GAGCAAAGGT CCTAGGAAAT TTAAGTAAGC CATTTACCAA CGCTCAGAAG 7000
AAAGAACTTG AAGAGCACTT GGAAATGAGC TGTGTCTCCC CAAGAAAGAG
GGAGAGAAAG AGGGGAGAGA TGTGGTGCAG ACCCTAGGGA GGAAGGAGTT 7100
CAGAAAAACC ATCCTCAGGG TGTTCTTGCT ACAAACCAAA AAATGCAGCA
TGGTGGTGGG GAGGATGACT CTGTCCTCCC TGACTTTTAG ATGAGCCCAA 7200
GGGAAAAGGC AAAGACAAAG CCCTTAAGAG CCAGAGGACT CACGAGGGCC
TGGGGCTGGT GAGAGTGGCG GGGAGAGAGG GCTCACCTTG GGAGAAGGAT 7300
GGTCAGTGTC TGGGGCTTTC CTGGTCATGT TCCAAATCAG GCTTGGCAGG
AGTCCTGCTG TGCAAATTGC GTTTGCTGAG CCCTGTCAGA GGTCTCCTGT 7400
GTCTCACATC TAGGGTGACC AGCATCCTGG CTTCCTCAGG ACTGTTCAGG
TTTTAGCACT GAACATCACA TGTCCTAGGG AACCCCTCAG TTTGGGCAAG 7500
CCCTGCCACA TCACACAATC ATATTAGTGC CCTCAGTATT CTTTGCAAAC
ATAAAACCAT AGACTCAGTA ATCCCATTAC TGGGTATATA CCCAAAGAAA 7600
TATAAATTAT TCTACTATAA AGACACATGC ACATATTTGT TTATTGCAGC
```

FIGURE 1C

```
ACTATTCACA ATAACAAAGT CATGGAACCA ACCCAGATGC CCATCAATGG    7700
TAGATTGGAT AAAGAAAATG TGGTACATAT ACACCATGGA ATACTATGCA
GCCATAACAA GGAATGAGAT CATATTCTTT GCAAGGACAT GGATGAAGCT    7800
GGAAGCCATC ATCCTCCACA AACTAACACA GGAACAGAAA ATCAAACACC
GCATGTTCTC ACTCATAAGT GGGAGTTGAA CAGTGAGAAT GCGTAGACGC    7900
AGGGAGGGGA ACAACACACA CCAGGGCTTG TGGCGGGGTG AGGGGTGAGG
GGAGGAACTT AGAGGATAGG TCAATAGGTG CAGCAAACCA CCATGGCATA    8000
TGTATCCCAG AACTTCAAGT AAATAATAAT AATAATAATT AATAATAATA
ATAATAATAA ATAAACCCAT AAAGCCATTT GAGAGATTCT TGGGGGATTC    8100
ATTGGACCAC TGAAAATCTA CAGTGAGAAA AGAATTGCCA TGTTGATGAA
ACAGGAAAAC TTTCCTTGTC CCCCTCACAG AGCATGTGAC AGCGGGAGGG    8200
GCTCACTTTC TCAGTGCGCC ACTGCTCAAA CCTCTAGGGG AGCATACAGA
CGGGCAGGTT GTGGGGCTCT GACCTCACCG GCAGTGTCTA GAGGTGGATG    8300
TTTACAGCTC CTGAAGCTCC AGTGGGCGTG GGTTATGGCC TTCTTTTAGT
TTTGCCCTCT ATAGTCAGCT TGTGTTAACC AGCTCAATTA CACCCTCTAC    8400
CTTGTCGCAA GGACAGAGGG CTTTCTGTAT CCTGGGGGCT TGCCTTGGTG
TACCAGAAGA ATCGAATCCC ACCTGGGCTT GGAGAATGAG TGCAAGGATT    8500
TATTGAGTGG ATGTAGCTCT CAGCAGATGG GGGAAGCCAG AAGGGGATGG
AATGGGAAGG GTTTCCCCTG GAGTCAGACC GCTCAGTGGC CCGGGCTCGG    8600
TGGCCCGGGC TCGGTGGCCT GGGCTCTCCT CCGACTGCCT CAGCCAAACT
CCGCGTTGTT CTGCTGGTCA GTGGCCTGCC GGTGCCTGTT GGTGAGTTCT    8700
TCTCAATGTC CAGCTGTCCT TGCGTCCCTC CGCTGATGTG CTCCTCCCGA
TGTCCAGCTA CCTGTGTGTC TGCCTGCTAG GGTCTTGGGG TTTTTATAGG    8800
CACATGATGG GGGCGTGGCA GGCCAGGGTG GTTTTGGGAA ATGAAACATT
TAGGCAGGAA AACAAAAATG CCTGTCCTCA CCTAGGTCCA TGGGCACAGG    8900
TCTGGGGGTG GAGCCCTCGC CAGGGACCAC ACCCTCTTCT ACCCAGCACT
TCCCTTCCCT ACTTCCATAT CATTTAAAGG GACCACGCCC TTCCCAGCTC    9000
TTCCCTTCTG TATCACTGAT GCCTTGCTCT GTGTTCTCTA AGTGGAATTA
TCACTGTGTG TATGTACAGG TGTGTGCATG TGTGTGCATG TACCTGTGCT    9100
TTTCTTTTGG AAAACTAGCA CATTACCTGG ATTTTGCATC TCAAGGATAA
TTCTGTAAGC AGGAACCCTT CCTCCTTTAG AAGGAAGTAA AGGAGAGGAA    9200
AATGCTGTAA AACTTACATA TTAATAATTT TTTACTCTAT CTCAAACACG
CATGCCTTTA ATCATAGTCT TAAGAGGAAG ATATCTAATT CATAACTTAC    9300
TGTATGTAGT CATCAAAGAA TATGAGAAAA AATTAACTGA AAATTTTTCT
TCTGGCTCTA GGAATTTGAA GCCTACATTA ATGCTTCTGG AGAACATGGA    9400
        [EXON 2: 9362..
ATTGTGGTTT TCTCTTTGGG ATCAATGGTC TCAGAAATTC CAGAGAAGAA
AGCTATGGCA ATTGCTGATG CTTTGGGCAA AATCCCTCAG ACAGTAAGAA    9500
                     ..9493]
GATTCTATAC CATGGCCTCA TATCTATTTT CACAGGAGCG CTAATCCCAG
          C    T
ACTTCCAGCT TCCAGATTAA TTCTCTTAAT TGGAACCTTA GATTTGGCTT    9600
TTCCCTGCCA CTTCCCAACT ATTAATCCAA AGGTTTTTTT TGTTGTTGTG
GTTGTTGTCA TTGTTTTCAA TTTGACTCTC AAATACTCTA TTAAACTATG    9700
ATCCACCACA CTCAGAAGTA TCATTTTCTC TAAGAGACTC AAAAGTGTAT
TAGGGAGAAT TTATTTAAAA ATAAAATAAA TGGGATATTG TTTCTTCATA    9800
TTAAATAGAA GTATTTCTCC AAAAAGCTGT TGGTTAGAAC ACTGAATTTA
TGTCTTACAT TTCTGCTCTT ATAGTTCTGC ATCCACTTGT TTCATTAAGC    9900
AAACTTTCCC TTAAAGTGCA GGAAAGTGAA AAAATCCTAA GTGCACAGCT
TGATAAATTA TCACAAATTC ACGTAGTGCA TACACCCTTG TAACTAAACC    10000
TCCAAAACAA GATGCCGGAA GTTGCCAGTC CTCAGAAGCC TTCACAGTTA
CTGATCCTCC CACTCTGTTA AAGACTGTTC CTTCAGAGGA CCCCTGTTTT    10100
                                            T   TC
```

FIGURE 1D

CTAGTTAGTA TAGCAGATTT GTTTTCTAAT CATATTATGT TCTTTCTTTA
C
CGTTCTGCTC TTTTTGCCCC TCCCAGGTCC TGTGGCGGTA CACTGGAACC 10200
[EXON 3: 10177..
CGACCATCGA ATCTTGCGAA CAACACGATA CTTGTTAAGT GGCTACCCCA
AAACGATCTG CTTGGTATGT TGGGCGGATT GGATGTATAG GTCAAACCAG 10300
..10264]
GGTCAAATTA AGAAAATGGC TTAAGCACAG CTATTCTAAA GGATTGTTGA
GCTTGAAAAT ATTATGGCCA ACATATCCTA CATTGCTTTT TATCTAGTGG 10400
GGTATCTCAA CCCACATTTT CTTCTGCAAA TTTCTGCAAG GGCATGTGAG
TAACACTGAG TCTTTGGAGT GTTTTCAGAA CCTAGATGTG TCCAGCTGTG 10500
AAACTCAGAG ATGTAACTGC TGACATCCTC CCTATTTTGC ATCTCAGGTC
[EXON 4: 10548..
ACCCGATGAC CCGTGCCTTT ATCACCCATG CTGGTTCCCA TGGTGTTTAT 10600
GAAAGCATAT GCAATGGCGT TCCCATGGTG ATGATGCCCT TGTTTGGTGA
TCAGATGGAC AATGCAAAGC GCATGGAGAC TAAGGGAGCT GGAGTGACCC 10700
TGAATGTTCT GGAAATGACT TCTGAAGATT TAGAAAATGC TCTAAAAGCA
GTCATCAATG ACAAAAGGTA AGAAAGAAGA TACAGAAGAA TACTTTGGTC 10800
..10767]
ATGGCATTCA TGATAAAATT GTTTCAAATA TGAAAACATT TACGTAGCAT
TTAATAGCGT TGTTTCAAAT ATAAAAACAA ATACATAAAA ATCTGGATTT 10900
TTATTTCTTC TTTTTTTTTT TTTTTTTTTT TTGAGATGGA GTCTTGCTCT
GTCACCTAGG CTGGAGTGCA GTGGTGCAAT CTTGGCTTAC TGCAACCTCC 11000
ACCTCCCACG TTCAAGCAGT TCTGCCTCAG CCTCCGTGTA GCTGGGATTA
CAGGTGTCCA CCACCACGCC CGGTTAATTT TTGTATTTTT TAGTAGAGAA 11100
AGGGTTTCAC CATGTTTGTC AGGCTGGTCT TGAACTCCTG ACTTCAGGTG
ATCCACCTGC CTCGGCCTGC CAAAGTGCTG AGATTACAGG CATGAGCCAG 11200
CGCGTCTGAC CTGGATTTAT AAATAAGATA ATTTAGAGGT TATTATTCAC
TTTATAAAAG GATTCTTTAG TTTCTATATA ATTTATCATA TAATTTATTT 11300
AGAATTTTAT TTCCCCCATT AGATTTAAAA CTCCAATTTA CATAAAAAGT
TGCCATAATA GACATCTGAT CCATAAGTTT CCTGCACAGA AAGAAATACT 11400
CCATTATAAG AAGCATAGTA TCTTTAAGAG AAAAACAACT CAAATGCTTA
GAAGTACAGC TTTTTGCAGC ACTGGAACCT GTGAGAAATT TTGTCCATGG 11500
AGTTTATGAA TGAAGGAGCT ATAAGATATC ACAGACAAAG TCTTAGAATA
AGAGCAAAGG AAAATTTGCT CAAATGTGGC CCTGAAAACG ATTCAAAGGG 11600
CAAATGATTT CTGGATTAAA GTTAGTATAT TACTGTCAAG CTCACTGGTA
ATAGGCTTAT TAGAACCTTA TGGGAAGAAG TGGTGGCCAG TGGTAGATTT 11700
CATCCGACAA TAGATACTGT GTGCATATGT GCGTGTGCGT TTGTGCATGT
GGCTGTGCTC ATGTGTGGGT GCACACGTGT GCATTCATAT GCGTGTGTGT 11800
GTGTGTGCGT GTGTTTATGA GAGTGTCCAT TGCTTTCTCC CATGGTTACC
TCCTTTAGAA AGAAGCAGCA GTCAGGAAGA CAGATGTGAA GAGCTGGAGC 11900
ATGTTCAGAT GAGAGGAGAC GGAACACGGG GACACACCAG CTTGAGCAAG
GGACAACAGG GGAGGACTGA TGACTGACTT CCCACCTTTG AGGTGCTAAT 12000
GTGTGTGTGG TGGCACTGGA TAAAAGATCA ATGTTGGCTA GGCACCATGG
CACACGCCTG TAGTCCCAGC CACTCTGGAG GCTAAGGCGG GAGGATTGCT 12100
TGAGCCCAGA AGTTGGAGGC TGCTATGAGC CGTGATCATG CCACTGCACT
CCAGCAACCT GGGCAACAGA GTGAGACCCT GTCTCAAAAA AAAAAAAAAA 12200
AATGAAAAGT CCACATAACC TGAGCATCAT GTGCCCAGAG CGTTGGGTGG
TGTGGTCCCA TTCCTTCCTT CCAGCGGCTT CTTCTGGCCA CCTCAATGTC 12300
AGGATGTCCT GCTCACATAT CAATACCATT AAAACCTGAC TTCTTTCCCT
GCACTGTTGA AGCTCCTTCT TGAGGCTCAC ATTATGGATA TAATTTTGAT 12400
TCTTTCTTCA GTGGTATAGA TAACTACTTG TAACCTAAGA ACAACTTGGT
GAAAGTCCTC TAATACATTA TTTTTTAAAA AAACACAAAT CAATGAGCTC 12500

FIGURE 1E

AACTTATTAA CTAACTTTCA TCTATTCATT TTTGAGCCAT CCCTGTCTGA
TTGTGAATCT CCATGATTCC AACACTCTGA GCTGGGGATA GTGCCTACAC 12600
AAAATAAAAA GAAGTGGAAA ATTTTCAAAC ATCAGTTTAT GCTGACAACC
AGGCCATAAT AGGTGCTCAA TTACTATTGA ATGAATGAAT GAAAGTTCTG 12700
GCCAGGTACG GTGGCTCATG CCTGTAGTCC CAACACTTTG GGAGGCCGAG
GCAGGTGGAT CACTTGAGGT TAGGAGTTCG AAACCAACCT GACCAACATG 12800
AAGAAACCTT ATCTCTACCA AAAAAATATA AAAAAATTAC CCAGGCATGG
TGGTGTATGC CTGTAATCCC AGCTATTTGG GAGGCTGAGG CAGGAAAATC 12900
ACTTGAACCT GAGAGGCGGA GGTTGCAGTG AGCTGAGATT GTGCCACTCC
ACTCCAGCCT GGGCGACAGA GTGAGACTCC GTCTTACTTA AAAAAAAAAA 13000
AAAGAAGGTT CCAAGAAAAT TCATCTTAAG GTTTATGTAA AAGGAAGATG
ATATTTAACA TGATTCATGG CCAAGTACTA ATATTACATT ATAATAATGT 13100
TTCCAAATAA CATTATAGAT ATGTTTAAAG ACAGTGTATT AGGCTGTTCT
TGCATTGCTG TAAAGAAATA CCCAAGACTG GGTAATTTAT AAAGAAAAGA 13200
GGTTTCATTG GCTCGTGTTT CTGCAGGCTG TACAGGAAGC TTAGTGCTGA
CATCACTTGG CTGCCGGGGG AACCTCAGGG AGCTTTTACT CATGGCAGAA 13300
GGCAATGCGG GAGCTTGCAT GTCACATGGC AAAAGCAGGA GCGAGAGAGA
GTTGGGGGGG AAGGTGCCAC ACACTTTTTA ATGACCGGCT CTCACAATAA 13400
CTCATGAAAA CTCACTATCA GGAAGACAGC ACTAAAGCAC AAGGGATCCG
ACCCCATGAT CCAAACACCT CCCACCAGGC CCCATCTCCA GCACTGGGGA 13500
TTACAATTCA ACATGAGATC TGAGTGTGGA CAAATATCCA AACTGTATCA
GTCAACAGCG ATCATAATTA GTCCTGAATA GGAGTGCCTT TTTTTTTCTT 13600
TCTTCTCCCT TTTCTTTTCT ACTTCCTCCT CCTTTTCCCT CTCCTCTTCA
ATCTCCTCTT CATTCCTGTA GCACCAAGGG TTGAAGCACC TAACCCGTTT 13700
TGGATTGAGA TGTTCTGATT GGGCAATGAA CACTGTCCAG AATAAACAGA
AATCCATTTT GCACTAAGTG GCTGCACAGA CCCTGCCTCA TGCTAAATCT 13800
AGCACCCAGA TAGTTTAATG TTTCAATGAC TGAATTACAA ATATATCATC
ACCTTGGATT TGGCACTTAC AAATGGCTGT TAATTTGGCC AGAGGTGGTT 13900
GTTTACAACT TCAAATAGGA GACTATTCAT AATTTCTGAC GTGACATTTT
CCTTTCTTTA TTTTACTGTA TGAAAATATA ATGAAATTTC TCACAAAATA 14000
TCACTAAAAA GAAAAGAAGA AGAGTAGGAA GCAAGGTTAA AATATTTCTA
AAATATAATT TTGGTCTTTC TTTTTCTCCC TTCCTTCCTC CGTCCCTCTC 14100
TCCTTTCCTC TCTCCCTCCC TCCCTCCCTC CCTTCCTCCT TTCCTTGCTT
CCTTCCCTCC TTCTCTTCCT TCTTTTTCAA GAGATCAATA ACATTTATTA 14200
AGAATAAGTT TCTTAATTAT AACCTTTCAG GTGATAATAG TAACACAGCC
TGGGCAACAC AATAAGACCT TGTTTCTACA AAAAATTTAA AAATTGGCCA 14300
GACATAGTGG TGCATGACTA ATTCCAGCTA CTCTGGAGGC TGAGGCAGGA
GGATGGCTTG AGCCCAGGAG TTGGAGGCTG CAGTTAGCCA TGCTTGTGCC 14400
ACTACACTCC AGCCCGGGCA ACAGGGCAAG ACTCTGTATC TAAAAACAAC
T
AACAACAACA ATAATAGAAA CAGGTTTCCT TTCCCAAGTT TGGAAAATCT 14500
GGTAGTCTTC TTAAGCAGCC ATGAGCATAA AGAGAGGATT GTTCATACCA
CAGGTGTTCC AGGCATAACG AAACTGTCTT TGTGTTTAGT TACAAGGAGA 14600
[EXON 5: 14590..
ACATCATGCG CCTCTCCAGC CTTCACAAGG ACCGCCCGGT GGAGCCGCTG
GACCTGGCCG TGTTCTGGGT GGAGTTTGTG ATGAGGCACA AGGGCGCGCC 14700
ACACCTGCGC CCCGCAGCCC ACGACCTCAC CTGGTACCAG TACCATTCCT
T
TGGACGTGAT TGGTTTCCTC TTGGCCGTCG TGCTGACAGT GGCCTTCATC 14800
T
ACCTTTAAAT GTTGTGCTTA TGGCTACCGG AAATGCTTGG GGAAAAAAGG
GCGAGTTAAG AAAGCCCACA AATCCAAGAC CCATTGAGAA GTGGGTGGGA 14900
..14887]
AATAAGGTAA AATTTTGAAC CATTCCCTAG TCATTTCCAA ACTTGAAAAC

FIGURE 1F

AGAATCAGTG TTAAATTCAT TTTATTCTTA TTAAGGAAAT ACTTTGCATA 15000
C
AATTAATCAG CCCCAGAGTG CTTTAAAAAA TTCTCTTAAA TAAAAATAAT
AGACTCGCTA GTCAGTAAAG ATATTTGAAT ATGTATCGTG CCCCCTCTGG 15100
TGTCTTTGAT CAGGATGACA TGTGCCATTT TTCAGAGGAC GTGCAGACAG
GCTGGCATTC TAGATTACTT TTCTTACTCT GAAACATGGC CTGTTTGGGA 15200
GTGCGGGATT CAAAGGTGGT CCCACGGCTG CCCCTACTGC AAATGGCAGT
TTTAATCTTA TCTTTTGGCT TCTGCAGATG GTTGCAATTG ATCCTTAACC 15300
AATAATGGTC AGTCCTCATC TCTGTCGTGC TTCATAGGTG CCACCTTGTG
TGTTTAAAGA AGGGAAGCTT TGTACCTTTA GAGTGTAGGT GAAATGAATG 15400
AATGGCTTGG AGTGCACTGA GAACAGCATA TGATTTCTTG CTTTGGGGAA
AAAGAATGAT GCTATGAAAT TGGTGGGTGG TGTATTTGAG AAGATAATCA 15500
TTGCTTATGT CAAATGGAGC TGAATTTGAT AAAAACCCAA AATACAGCTA
TGAAGTGCTG GGCAAGTTTA CTTTTTTTCT GATGTTTCCT ACAACTAAAA 15600
ATAAATTAAT AAATTTATAT AAATTCTATT TAAGTGTTTT CACTGGTGTC
GCATTTATTT CTTGTTAAGT TGCATTTTCT AATTACAAAA GTAATGCATG 15700
ATTATGACAG AAAGTTTGGA AAATATAGAG GTTCACACAC ACACGCCTTC
ATTGCGTGTG CATGCATAAA TGCATGAGAA AAGAAAAATA ACCAGTAATC 15800
ACATCGCCCA GAAATAACCC CAGTTACAAT TGTGGCAAAT ACACATACTT
ATAAATATTG CAGATATATT AAGTATACCT AGTATTTGCT AACACTCTTT 15900
CTTCTACTCT GTCATGAAGA TTCTCCCAAG GTGTTTTTGT ATAATATTTA
ATTCATTTTC AGTGGCCAAG CAGTATTCTA CTTCATGGAT ATACCAGGAT 16000
TTATTTAACC ATAACTTCTG GTTGGATTCA CTCTTATTAT TTTGTTTAAT
TAAAAAAAAA AGACCTCGGC TGGGCACAGT GGCTCATGCC TGTAATCCCA 16100
GCACTTTGGG AGGCCGAGGT GGGTGGATCA CCTAAGATCG GGAGTTTGAG
ACCAGCCTGG CCAACATGGC AAAACCCCGT CTCTACTAAA AATACAGAAA 16200
ATTAGCCGGG TGTGGTTGCC AGCACCTGTA ATTCCAGCTA ATTGGGAGGC
TGAGGCAGGA GAATTGCTTG AACCGGGGTC AGGGGGTTCG GAGGTCGGAG 16300
GTTGCAGTGA GTCCGGATCA TGCCACTGCA TTCCAGCCTG GGTGACACAG
CCAGACTCTG TCTCAAAAAC AACAACAACA ACAAAACAAC AACAACAACA 16400
ACAACAAAAA TCTCACTGGA CATCCTAGTA GCTAAGGCTT TCCACATATT
CATGATTACT TCTGTTGGAA AGTGCTTTAC AACAAATTGC TAGTTGTCTC 16500
AGTCTGGGTT CCCCTGAGAT GAGGATTCAA GGGCCAGGAG TTTATTTAGG
AAGTAAAGGA AACACTGATA GAGGAGTGGC AGAGTGAGAA GGGGTGATGG 16600
TCATCCACAG CTGGCTCTCT TGTGGTCAAT CGGAGCTTAA TCCTGCTGGG
TGACTCTGGG AGCCAGTGGA GAAAAGACAC CCCAGACTTA TCCCAATGAG 16700
GAACACGGCT GTTGGGTGCC TGAGTACTTG CCTCGTCAGG GATTGAAACG
TACTCCCAGG TAGTAGTAAT TTCTCTGCCC TTCCATTAGG CCACAAAGGG 16800
GGCTCTGACA GAGAGAGCTG ACGAGAAAAA ACACACGCCC TTGTCACTGA
AGAGGTACAC AGGGGATCTG TGTGGGGCAC CACCTGCACT GCTACCCTGG 16900
ACAAATAGCT TAAGAAATCC CCACACTGCA TCCCCAAACT TACTATCAGC
GTGTGAGGGA GACAGGTTCC CACACCCTCA TTAGCACAAA GTACTATCTT 17000
GAAAAAGAAA GCCTGTCAGT TTGATAGGAG AAAAGCAGGA TCTTGTTTAC
AATGTGCTTT TATTATTGTT ATTATTAGAG ATTGTATTTC TTTTCAAGCT 17100
GATGAGCCGT CTGTGTTTAT TTTTTGGAGG ATACCCTTTG CCCACTTTCC
TATTGGAGTG TATTACCCTG AGGATTTGGT AAGAGTGCTT ATTGCATTCA 17200
CCAGAATGTC CTTTTTGTCA TTTACTGTAT TTTCTCTACT TTTTTTTTTT
TGCCTTGTTT TACTTTTTTT GTTTTGTATT ACAAGCAGAA GTTTTAAATT 17300
TGTAAGCTTC AAATTGGAGC TGGGGTGGTG CAGAGCGAAG ATTTCAGCTG
GTTCCCTGAC CCCAGCTCCA TCTCCTTCCC TAGGCAGTGG CTGGAACACA 17400
TTCTGTCCAC TATTTCCCTC TCTACATCCT TGAGGCTGTG CAGTCACCCC
TCAACTACGT TCACCCTCCT TCAAAGCCCT TCCTGGTCCA CCCGGGGACC 17500
ATCTCCCGGC CTCACTGCCC CTAGCTCCTT GACGCCCCAA CCTCTCTCAG
GGACCCCAAG TTGCCATGAC CTCCAGCCAG CTCATGTTCA TTTGCACCTT 17600

FIGURE 1G

CGTGTCTGCA GCACTGAGGC ACTCTTGTTT ACAAGTGAGA GAACCCAACT
CGGGATACCT TAAGCATAAA CAGTATTTTT GTAAGGAGAC AGGCTTCTGA 17700
CGACGCGAGG CTCATAGCCA GGCCTGCGCT GGGTGGAGCC TCCCCTTCTC
ACTCCTGTCC CTGTTGGGTC AGAGTGCCAG CTTTCCTCTT CCCTCTCCTC 17800
CGGGTCTTTT CGGCCCCTCA GTCCCCATAT TCCTCTGCCC TAGCTCCCAA
GATCCCACAA GAGACAGACT GGATTCTCTC TGGCCTGGAG TGCCACCTTC 17900
CTGAAAGTCA GAATCTGATT GGTCCAGCTG GATCAGGTGT CCTCTCCCTG
TCCAATCATC AATGCCGAGA GGGATTACGG AGAGAAAAAC ATGGCTCCCA 18000
CCATCCCATT GCTGTGGCTG CTTGGGCCAC GGGAGAGGGA ACCTTGTGAG
CCGGGCAGAA TCCACCCTGT AGAGACTGCC TCTGGGTGAG TCATATGGTT 18100
TGGCTGTGTC CCCACCCAAA TCTCATCTTG AATTGTAATT CCCATTACCC
CCATGTGTCA TTAGAGGGAC CTGGTGGGAA GTGATTTGAA TTATGGGGGC 18200
AGTTATCTCC ATGCTATTTG TGTGATAGTG AGTTCTCACA AGATCTGATG
GTTTTATAGG GGGCTTTTCC CCCTCTTGCT CATACTTCCT CTTGCCTGCC 18300
ACCATGTAAG ATGTGCCCTT GCTCCTCCTT CACCTTCTGC CATGATTGTG
AGGCCTCCCC AGCCATGTGG AACTGTGAGT CCATGAAACC TCTTTTTCTT 18400
TATAAATTAC CCAGTCCTGG GTATTTCTTC ATAGCAGTAT GAAAATGGAC
TAATACAGTC AGCTTCTGCA CAATATTTCC ATTTTCCCAC ATTATGTCTT 18500
GGGCCTTTTG TGTATTTAAG CTCACAGGAT GCTACGAATA AAGCGTTTTC
TTATTTCCTG GGTAGTTCCC ATAGAAGTAG TGGTGCAACG TGCCATAGAG 18600
TGACAGCACC TAAGAGAAGC TGATTTTGTG AGTGGATTGT GAGTTCAATA
TTGTTGTCAT AATCAGAAAA AAATGTATTT ACTTTTTTTT TTTTTTTTTT 18700
TGAGACGGAG TCTCACTCTG TTGCCCAGGC TGGAGTGCAG TGGTGTGATC
TTGGCTCACT GCAACCTCCG CCTCCTGGGT TCAAGTGATT CTTCCTACCT 18800
CACCCTCCTG AGTAGCTGGG ATTACAGGCA CATGCCACCC CACCACACCC
GGCTAATTTT TGTATTTTTT AGTAGAGACA GCGTTTC 18887

FIGURE 1H

POLYMORPHISMS IN THE CODING SEQUENCE OF UGT1A1

ATGGCTGTGG AGTCCCAGGG CGGACGCCCA CTTGTCCTGG GCCTGCTGCT
GTGTGTGCTG GGCCCAGTGG TGTCCCATGC TGGGAAGATA CTGTTGATCC 100
CAGTGGATGG CAGCCACTGG CTGAGCATGC TTGGGGCCAT CCAGCAGCTG
T
CAGCAGAGGG GACATGAAAT AGTTGTCCTA GCACCTGACG CCTCGTTGTA 200
CATCAGAGAC GGAGCATTTT ACACCTTGAA GACGTACCCT GTGCCATTCC
A
AAAGGGAGGA TGTGAAAGAG TCTTTTGTTA GTCTCGGGCA TAATGTTTTT 300
GAGAATGATT CTTTCCTGCA GCGTGTGATC AAAACATACA AGAAAATAAA
AAAGGACTCT GCTATGCTTT TGTCTGGCTG TTCCCACTTA CTGCACAACA 400
AGGAGCTCAT GGCCTCCCTG GCAGAAAGCA GCTTTGATGT CATGCTGACG
GACCCTTTCC TTCCTTGCAG CCCCATCGTG GCCCAGTACC TGTCTCTGCC 500
CACTGTATTC TTCTTGCATG CACTGCCATG CAGCCTGGAA TTTGAGGCTA
G
CCCAGTGCCC CAACCCATTC TCCTACGTGC CCAGGCCTCT CTCCTCTCAT 600
TCAGATCACA TGACCTTCCT GCAGCGGGTG AAGAACATGC TCATTGCCTT
TTCACAGAAC TTTCTGTGCG ACGTGGTTTA TTCCCCGTAT GCAACCCTTG 700
CCTCAGAATT CCTTCAGAGA GAGGTGACTG TCCAGGACCT ATTGAGCTCT
GCATCTGTCT GGCTGTTTAG AAGTGACTTT GTGAAGGATT ACCCTAGGCC 800
CATCATGCCC AATATGGTTT TTGTTGGTGG AATCAACTGC CTTCACCAAA
ATCCACTATC CCAGGAATTT GAAGCCTACA TTAATGCTTC TGGAGAACAT 900
GGAATTGTGG TTTTCTCTTT GGGATCAATG GTCTCAGAAA TTCCAGAGAA
GAAAGCTATG GCAATTGCTG ATGCTTTGGG CAAAATCCCT CAGACAGTCC 1000
TGTGGCGGTA CACTGGAACC CGACCATCGA ATCTTGCGAA CAACACGATA
CTTGTTAAGT GGCTACCCCA AAACGATCTG CTTGGTCACC CGATGACCCG 1100
TGCCTTTATC ACCCATGCTG GTTCCCATGG TGTTTATGAA AGCATATGCA
ATGGCGTTCC CATGGTGATG ATGCCCTTGT TTGGTGATCA GATGGACAAT 1200
GCAAAGCGCA TGGAGACTAA GGGAGCTGGA GTGACCCTGA ATGTTCTGGA
AATGACTTCT GAAGATTTAG AAAATGCTCT AAAAGCAGTC ATCAATGACA 1300
AAAGTTACAA GGAGAACATC ATGCGCCTCT CCAGCCTTCA CAAGGACCGC
CCGGTGGAGC CGCTGGACCT GGCCGTGTTC TGGGTGGAGT TTGTGATGAG 1400
GCACAAGGGC GCGCCACACC TGCGCCCCGC AGCCCACGAC CTCACCTGGT
T
ACCAGTACCA TTCCTTGGAC GTGATTGGTT TCCTCTTGGC CGTCGTGCTG 1500
T
ACAGTGGCCT TCATCACCTT TAAATGTTGT GCTTATGGCT ACCGGAAATG
CTTGGGGAAA AAAGGGCGAG TTAAGAAAGC CCACAAATCC AAGACCCATT 1600
GA 1602

FIGURE 2

ISOFORMS OF THE UGT1A1 PROTEIN

```
MAVESQGGRP LVLGLLLCVL GPVVSHAGKI LLIPVDGSHW LSMLGAIQQL
QQRGHEIVVL APDASLYIRD GAFYTLKTYP VPFQREDVKE SFVSLGHNVF    100
                      R
ENDSFLQRVI KTYKKIKKDS AMLLSGCSHL LHNKELMASL AESSFDVMLT
DPFLPCSPIV AQYLSLPTVF FLHALPCSLE FEATQCPNPF SYVPRPLSSH    200
SDHMTFLQRV KNMLIAFSQN FLCDVVYSPY ATLASEFLQR EVTVQDLLSS
ASVWLFRSDF VKDYPRPIMP NMVFVGGINC LHQNPLSQEF EAYINASGEH    300
GIVVFSLGSM VSEIPEKKAM AIADALGKIP QTVLWRYTGT RPSNLANNTI
LVKWLPQNDL LGHPMTRAFI THAGSHGVYE SICNGVPMVM MPLFGDQMDN    400
AKRMETKGAG VTLNVLEMTS EDLENALKAV INDKSYKENI MRLSSLHKDR
PVEPLDLAVF WVEFVMRHKG APHLRPAAHD LTWYQYHSLD VIGFLLAVVL    500
TVAFITFKCC AYGYRKCLGK KGRVKKAHKS KTH                      533
```

FIGURE 3

SEQUENCE LISTING

<110> Genaissance Pharmaceuticals, Inc.
Koshy, Beena
Rounds, Eileen
Chew, Anne
Choi, Julie Y.

<120> HAPLOTYPES OF THE UGT1A1 GENE

<130> MWH-0236PCT UGT1A1

<140> TBA
<141> 2001-04-13

<150> 60/197,514
<151> 2000-04-18

<160> 74

<170> PatentIn Ver. 2.1

<210> 1
<211> 18887
<212> DNA
<213> Homo sapiens

<400> 1
gaattcaagg gattcaagga aggtggcttt gcttcccggg agggtcctgt agatgatcta 60
cagggcactg gacatgttta tgttgctcct ttagtaataa gcctgtcatt ctgatttgat 120
gaaaggagat gaaaggagct ggtagtgtgt ctgatggtgg cctactaact tatgtcttca 180
gcttaaaaag aaagtagctt caaaagggtt ccagaaacac tttccatgga cgtgtcactc 240
tttagcagcc cccaaagcaa gaccatcata ttgctgccct gctgtgtgat ttctcagccc 300
ctagagcacc atcccctgta attgcctggt catgagtttg tctctgtcta cctgacccct 360
cctttcaggc aaggaccatt tctaacttga ctttctgggc ctagttccta gcatagtgac 420
tgccatccag tagggctcac acgttccata aatatttggc agatgaggga attagcaatg 480
ggttctgctt tggtttcaga gcagatatta attggattgc ttagtagtgg ttctctgttg 540
taattcatga gcatgaatgt ggattgccca ctattcagat tagtaagtat ttcttggtca 600
agggcagagc tgtggccaca aaccatccag gtacacagca gaagcagcct caaaaagctt 660
ggaagctctg catgatgcag gaaagtcata aaatcattac agtggtgact tatgtgttta 720
tagccccttt actgtctata atctgcaaat gaactcacac agcattggga ctttggaaga 780
attatcaccc ttaaggttta aattaaactg tgaatttcag aatttctaat aaggacacaa 840
caaagagtga aagcattgct atgtctattc tgcttgccca gaatcttggt cctaaaaaat 900
gaagagtgtt tgggtgtggg gaggagcttc agtgtgcatg tgcatgcaaa gtacctactc 960
taaggagaag aatgagaggg taccctaatt acctgttaat atgtcccata ggacaccaaa 1020
actctagtta gctgtttctc tatgatcctc taagcacatc cccaagtatg gctggccagt 1080
gatgtgtatg gttcaaatgt tgggatctgt gcagttatct tggaattgta tagtacagca 1140
gtatatcccc cccaaaaaga gtgtaatact tccaattctg gctgcacaat acttgcccca 1200

tagtccatgg tcaataaata caaatttgag ttgtttttgc tcatctttcc cttttgactt 1260
caaatcagtc atcagaattt ccccaaatgc ctttcccctg gatcttgggc cagtggaatg 1320
agtacaattt aacttaattg aatttgctta tctatttggt ttcctgttgt gaacaaaagt 1380
tctctgaaaa ggaatttgga agaaagagac tttgttctag tgaacagttt gcaaaccagg 1440
gagttacagc ctctggtacg caatgaaggt gagttccaca gaacacaagg caggcaggtt 1500
tcacggcaaa aagttccttc ccaggttccc aatcaggtcc atttatgcaa atgaaggatg 1560
gaaacttgct tagttcttat tggtcactgc agctgcattc tgattggttg atgaagctga 1620
gccctgagtg gctgaggtgg gtgagcttta attggttggt tcaggtgagc gctgaaaatc 1680
tcaactataa aaaggtacag gttttcagga tactcagagt aaccgtgtga cctgtagtaa 1740
gcaaagggcc agttggctct attttaaatc caggcccagt tagccactca agatctatct 1800
tacaggactg gctctttcag gttcacacta ataaaggcct gtccttgggg aagacttctg 1860
ttcacatgcg ctccagtgaa tttccctttc tggtcattct ctaccccagc acgcccccca 1920
cccccgaccc gccccaccca cccacctgtt catttccttc ttagcatgct tcacgatttc 1980
taagttcctg ctcatgtgtt taaattgtga gtctggctca cctcatggcg cgtgctcgtg 2040
tggtgggctc tgctgcagcc tcaagacccc acactgtgct ggactcaata aatattgttg 2100
gacgaaggaa tgaaacacat gatacaagtg agcaggcagt accgggggag ctgtggagtg 2160
ggcactctta caggtttcca tggcgaaagc gggggtacag ttgtgttctt ttctttctaa 2220
aaggctttct aaaaagcctt ctgtttaatt tctggaaaag aagcctaact tgttcactac 2280
atagtcgtcc ttcttcctct ctggtaacac ttgttggtct gtggaaatac taatttaatg 2340
gatcctgagg ttctggaagt actttgctgt gttcactcaa gaatgtgatt tgagtatgaa 2400
attccagcca gttcaactgt tgttgcctat taagaaacct aataaagctc caccttcttt 2460
atctctgaaa gtgaactccc tgctaccttt gtggactgac agctttttat agtcacgtga 2520
cacagtcaaa cattaacttg gtgtatcgat tggtttttgc catatatata tatataagta 2580
ggagagggcg aacctctggc aggagcaaag gcgccatggc tgtggagtcc cagggcggac 2640
gcccacttgt cctgggcctg ctgctgtgtg tgctgggccc agtggtgtcc catgctggga 2700
agatactgtt gatcccagtg gatggcagcc actggctgag catgcttggg gccatccagc 2760
agctgcagca gaggggacat gaaatagttg tcctagcacc tgacgcctcg ttgtacatca 2820
gagacggagc attttacacc ttgaagacgt accctgtgcc attccaaagg gaggatgtga 2880
aagagtcttt tgttagtctc gggcataatg tttttgagaa tgattctttc ctgcagcgtg 2940
tgatcaaaac atacaagaaa ataaaaaagg actctgctat gcttttgtct ggctgttccc 3000
acttactgca caacaaggag ctcatggcct ccctggcaga aagcagcttt gatgtcatgc 3060
tgacggaccc tttccttcct tgcagcccca tcgtggccca gtacctgtct ctgcccactg 3120
tattcttctt gcatgcactg ccatgcagcc tggaatttga ggctacccag tgccccaacc 3180
cattctccta cgtgcccagg cctctctcct ctcattcaga tcacatgacc ttcctgcagc 3240
gggtgaagaa catgctcatt gccttttcac agaactttct gtgcgacgtg gtttattccc 3300
cgtatgcaac ccttgcctca gaattccttc agagagaggt gactgtccag gacctattga 3360
gctctgcatc tgtctggctg tttagaagtg actttgtgaa ggattaccct aggcccatca 3420
tgcccaatat ggtttttgtt ggtggaatca actgccttca ccaaaatcca ctatcccagg 3480
tgtgtattgg agtgggactt ttacatgcgt atattctttc agatgtatta ctttggatcg 3540
attaactagc cccagatata tgctgagcaa gcattctgag ataatttaaa atgccctctt 3600
ttgttaattt ttgactccta ggtttgagtc tgtctttggc atcatcttct ggatgatttc 3660
ttggtatctg agatttcggg aaagcattcc ttggacattt tactctgtgt gctccagtgg 3720
atagtaatca attagaaaca acaagctgtt aaatgccata ggcacagaat gctgggtttg 3780
gggcaccctg cagaaaactc agttgaagcc tgcaccttgc cctggattca gtcaggcagg 3840
caatgttcag gactgatgaa atcattcttt gatgatgata gatcctggaa atgaaagttg 3900
cctttgtgac cctggttaaa gctccagttt ctaaatattc tgataagaag ctaaatcctg 3960
cagtccgttc tcttctaatg agtgaatcac cagacagtca ggttctgaca tgatacagaa 4020
aggttgtagg tttcattctc aagctattag gtttattttt cccctacaga gtttgaagta 4080

tgcaaaaagt agcattcaca tcctcatcga aatctcagca gaggatagaa aagaacagga 4140
gaggctcctt cagatggagc gttagggaat tactctttga ggaggtgaca tttcagagag 4200
cgttcattca cttatcctgc aaagattggc tgaggatcta ctggcagccc aggcacttcc 4260
caggtgctgc gtctggctcc cattaagggg actgatatca ccttcggagg tgaccttatt 4320
tccactatac ctccaatgtg atttgtattt tattttttt aattttctgt gcattttcct 4380
tcatagcaca tcaaatatgg cagccatttc acttagatag ttgttgattg tccgcttcac 4440
atcatgagcc atgtggggac ctgtgtgact ttgcattaat cacatccact gtatgcggcg 4500
tcctcaacac ctgccaatgg gtctgcatgt atttggcgcc ccataaatct cagcacctaa 4560
ggcacagaat aggcacccac cgaatatgtg ttacattaat gaatgagaag aaaggtgcca 4620
accgaggtct agttaatggg tcgagagtaa tccacaatag ctctttttag ttctttgtac 4680
tccagctatt acataccaat atgtatatag aaacatatgt aaaatttttt ggttgctttt 4740
tctacaaaat agagtaacag tgtattccca ctgcccactt accgataatg tcatggatat 4800
cactccagtt ttaaatgcta ttacttttta aactatgaaa tagtatttca tggtacttgt 4860
gtaccacagt gtattctgct ggagatctag tctagttccc cacagaggaa cattacaatt 4920
tgtattccag gagttttgtt gttgtgacct caaacacttc ctttaaaaag ataagctatt 4980
ttgtagttta aaaaacattt gttctgtttc tttctcattc atcttttctt aagtatttta 5040
cacggttttt tttttttggt cactactgtg aatgtgttat ttttttgcat ttctatctct 5100
agctgattat ctactcatta ctcagctatc tcatcaaaat attgattttc ataataaaaa 5160
ataataggca gtcatttgct gataaagaaa ttttggtttc ttctcttata aattccatgc 5220
caaatatcag ggctattgaa tttattagaa tctctaaaaa cagttgaata attctggcaa 5280
taggaaagat gcccgtcttg ctgctatttt agtggaaatt gattatcatt tcattatttt 5340
gcattatgtt agccattgtt ttctgaacag gctttattga tttagataat ttccttcttt 5400
gcgtgaggat gtttgtagga gaggcaccga actttatcag ctgcctttct ggcatttatt 5460
gatataacca taaaagtcta agtggtgaac tgtgttgact acatatttgt tgttgccttg 5520
tttggtgcag tcaggcttag gtgtgaaaat atgtttttaa attgtacctt ttagtaacct 5580
gttttgtctt gttgcatgtt ttaatctgaa attccacttt ttggatatta atattaccac 5640
ttctgtatta tttttgttta catttcccta gcacatcttt agtactcctt tgtcttcaag 5700
ctttcttcct ttttaaacaa catggcactg gtatttttaa tccagtcagg cagttgcttt 5760
aataagtgca ttttgcctat ttgaatctaa caattaatag atttgattgt aactctctca 5820
gtttacttta tgtttagttg actttgccat tctccttttt ccggatttct actggttggt 5880
caagttactg ttcttatttt ctctttcttc ctttgttaac taaaaatgcc actctgcact 5940
accattcctc ttgtgttgat ggtcctattc tcaatactct tgataaaact cctgaacttt 6000
aagaataaag ataaaacttt tattgcacaa agaagtccat agagaaagca caacctggca 6060
ttggcgtgtc tttggtgtgt ctgaaggaaa agagatagtg gaacaacatt gggagaaaag 6120
gaatgaaact caagaattcc aagatgttcc tcccctgcca gggtaagata gcagtggttc 6180
acagacaatc gcaatgctgg gtctgagaaa aataactaaa cagaagatta gtgaggacca 6240
aggcttcgag atggccagga gaggaaagct tgggagcagg gaaggttgag atatatgtgg 6300
gttactggga atgcgtgatg gtgaagtcac agatgaccca catggtgtct aagtgctaaa 6360
gaagaattct gggaaaatga aatgcatttg ggaagggaaa atctaattaa aagcctaaac 6420
taaaaataca aaattcttgg taaagtttag gagttatgtt aaatgtctca ttttggctgg 6480
tgaagtctca tcagaacagg gaaattctct cattcagggg catctcatct tttctttgaa 6540
gggaatcaat ggtgggggat tggagtgtta ttttcagtta atatgttgct tcactctttg 6600
gtcattccgg taactgtgaa gtcagggtga agtttaaggg aagctttgcc aagtagggga 6660
tggacttcac ctttattgag cctcatagta gctggctcag gtaggagttg gccgtgatga 6720
caacttctct gcagtttgcc ctgcgtgaat ctccagatga acttttgtgc catttaaact 6780
ttcgtgatct cctgctattt aacttcgaat gtttatggac ctgtgggttc aattttgtgt 6840
gaatcacatc ctgctgattg ctgagtgggc gtgtgggagg gtgtgcctgg aggagaactt 6900
agactcggcc ttttccagat gagcttcagt gtaagagtgg gtttcatgaa gagcaaaggt 6960

```
cctaggaaat ttaagtaagc catttaccaa cgctcagaag aaagaacttg aagagcactt 7020
ggaaatgagc tgtgtctccc caagaaagag ggagagaaag aggggagaga tgtggtgcag 7080
accctaggga ggaaggagtt cagaaaaacc atcctcaggg tgttcttgct acaaaccaaa 7140
aaatgcagca tggtggtggg gaggatgact ctgtcctccc tgacttttag atgagcccaa 7200
gggaaaaggc aaagacaaag cccttaagag ccagaggact cacgagggcc tggggctggt 7260
gagagtggcg gggagagagg gctcaccttg ggagaaggat ggtcagtgtc tggggctttc 7320
ctggtcatgt tccaaatcag gcttggcagg agtcctgctg tgcaaattgc gtttgctgag 7380
ccctgtcaga ggtctcctgt gtctcacatc tagggtgacc agcatcctgg cttcctcagg 7440
actgttcagg ttttagcact gaacatcaca tgtcctaggg aacccctcag tttgggcaag 7500
ccctgccaca tcacacaatc atattagtgc cctcagtatt ctttgcaaac ataaaaccat 7560
agactcagta atcccattac tgggtatata cccaaagaaa tataaattat tctactataa 7620
agacacatgc acatatttgt ttattgcagc actattcaca ataacaaagt catggaacca 7680
acccagatgc ccatcaatgg tagattggat aaagaaaatg tggtacatat acaccatgga 7740
atactatgca gccataacaa ggaatgagat catattcttt gcaaggacat ggatgaagct 7800
ggaagccatc atcctccaca aactaacaca ggaacagaaa atcaaacacc gcatgttctc 7860
actcataagt gggagttgaa cagtgagaat gcgtagacgc agggagggga acaacacaca 7920
ccagggcttg tggcggggtg aggggtgagg ggaggaactt agaggatagg tcaataggtg 7980
cagcaaacca ccatggcata tgtatcccag aacttcaagt aaataataat aataataatt 8040
aataataata ataataataa ataaacccat aaagccattt gagagattct tggggggattc 8100
attggaccac tgaaaatcta cagtgagaaa agaattgcca tgttgatgaa acaggaaaac 8160
tttccttgtc cccctcacag agcatgtgac agcgggaggg gctcactttc tcagtgcgcc 8220
actgctcaaa cctctagggg agcatacaga cgggcaggtt gtggggctct gacctcaccg 8280
gcagtgtcta gaggtggatg tttacagctc ctgaagctcc agtgggcgtg ggttatggcc 8340
ttcttttagt tttgccctct atagtcagct tgtgttaacc agctcaatta caccctctac 8400
cttgtcgcaa ggacagaggg ctttctgtat cctgggggct tgccttggtg taccagaaga 8460
atcgaatccc acctgggctt ggagaatgag tgcaaggatt tattgagtgg atgtagctct 8520
cagcagatgg gggaagccag aaggggatgg aatgggaagg gtttcccctg gagtcagacc 8580
gctcagtggc ccgggctcgg tggcccgggc tcggtggcct gggctctcct ccgactgcct 8640
cagccaaact ccgcgttgtt ctgctggtca gtggcctgcc ggtgcctgtt ggtgagttct 8700
tctcaatgtc cagctgtcct tgcgtccctc cgctgatgtg ctcctcccga tgtccagcta 8760
cctgtgtgtc tgcctgctag ggtcttgggg tttttatagg cacatgatgg gggcgtggca 8820
ggccagggtg gttttgggaa atgaaacatt taggcaggaa aacaaaaatg cctgtcctca 8880
cctaggtcca tgggcacagg tctgggggtg gagccctcgc cagggaccac accctcttct 8940
acccagcact tcccttccct acttccatat catttaaagg gaccacgccc ttcccagctc 9000
ttcccttctg tatcactgat gccttgctct gtgttctcta agtggaatta tcactgtgtg 9060
tatgtacagg tgtgtgcatg tgtgtgcatg tacctgtgct tttcttttgg aaaactagca 9120
cattacctgg attttgcatc tcaaggataa ttctgtaagc aggaaccctt cctcctttag 9180
aaggaagtaa aggagaggaa aatgctgtaa aacttacata ttaataattt tttactctat 9240
ctcaaacacg catgccttta atcatagtct taagaggaag atatctaatt cataacttac 9300
tgtatgtagt catcaaagaa tatgagaaaa aattaactga aaatttttct tctggctcta 9360
ggaatttgaa gcctacatta atgcttctgg agaacatgga attgtggttt tctctttggg 9420
atcaatggtc tcagaaattc cagagaagaa agctatggca attgctgatg ctttgggcaa 9480
aatccctcag acagtaagaa gattctatac catggcctca tatctatttt cacaggagcg 9540
ctaatcccag acttccagct tccagattaa ttctcttaat tggaacctta gatttggctt 9600
ttccctgcca cttcccaact attaatccaa aggttttttt tgttgttgtg gttgttgtca 9660
ttgttttcaa tttgactctc aaatactcta ttaaactatg atccaccaca ctcagaagta 9720
tcattttctc taagagactc aaaagtgtat tagggagaat ttatttaaaa ataaaataaa 9780
tgggatattg tttcttcata ttaaatagaa gtatttctcc aaaaagctgt tggttagaac 9840
```

actgaattta tgtcttacat ttctgctctt atagttctgc atccacttgt ttcattaagc 9900
aaactttccc ttaaagtgca ggaaagtgaa aaaatcctaa gtgcacagct tgataaatta 9960
tcacaaattc acgtagtgca tacacccttg taactaaacc tccaaaacaa gatgccggaa 10020
gttgccagtc ctcagaagcc ttcacagtta ctgatcctcc cactctgtta aagactgttc 10080
cttcagagga cccctgtttt ctagttagta tagcagattt gttttctaat catattatgt 10140
tctttcttta cgttctgctc ttttgcccc tcccaggtcc tgtggcggta cactggaacc 10200
cgaccatcga atcttgcgaa caacacgata cttgttaagt ggctacccca aaacgatctg 10260
cttggtatgt tgggcggatt ggatgtatag gtcaaaccag ggtcaaatta agaaaatggc 10320
ttaagcacag ctattctaaa ggattgttga gcttgaaaat attatggcca acatatccta 10380
cattgctttt tatctagtgg ggtatctcaa cccacatttt cttctgcaaa tttctgcaag 10440
ggcatgtgag taacactgag tctttggagt gttttcagaa cctagatgtg tccagctgtg 10500
aaactcagag atgtaactgc tgacatcctc cctattttgc atctcaggtc acccgatgac 10560
ccgtgccttt atcacccatg ctggttccca tggtgtttat gaaagcatat gcaatggcgt 10620
tcccatggtg atgatgccct tgtttggtga tcagatggac aatgcaaagc gcatggagac 10680
taagggagct ggagtgaccc tgaatgttct ggaaatgact tctgaagatt tagaaaatgc 10740
tctaaaagca gtcatcaatg acaaaaggta agaaagaaga tacagaagaa tactttggtc 10800
atggcattca tgataaaatt gtttcaaata tgaaaacatt tacgtagcat ttaatagcgt 10860
tgtttcaaat ataaaaacaa atacataaaa atctggattt ttatttcttc tttttttttt 10920
tttttttttt ttgagatgga gtcttgctct gtcacctagg ctggagtgca gtggtgcaat 10980
cttggcttac tgcaacctcc acctcccacg ttcaagcagt tctgcctcag cctccgtgta 11040
gctgggatta caggtgtcca ccaccacgcc cggttaattt ttgtattttt tagtagagaa 11100
agggtttcac catgtttgtc aggctggtct tgaactcctg acttcaggtg atccacctgc 11160
ctcggcctgc caaagtgctg agattacagg catgagccag cgcgtctgac ctggatttat 11220
aaataagata atttagaggt tattattcac tttataaaag gattctttag tttctatata 11280
atttatcata taatttattt agaattttat ttcccccatt agatttaaaa ctccaattta 11340
cataaaaagt tgccataata gacatctgat ccataagttt cctgcacaga aagaatatact 11400
ccattataag aagcatagta tctttaagag aaaaacaact caaatgctta gaagtacagc 11460
tttttgcagc actggaacct gtgagaaatt ttgtccatgg agtttatgaa tgaaggagct 11520
ataagatatc acagacaaag tcttagaata agagcaaagg aaaatttgct caaatgtggc 11580
cctgaaaacg attcaaaggg caaatgattt ctggattaaa gttagtatat tactgtcaag 11640
ctcactggta ataggcttat tagaacctta tgggaagaag tggtggccag tggtagattt 11700
catccgacaa tagatactgt gtgcatatgt gcgtgtgcgt ttgtgcatgt ggctgtgctc 11760
atgtgtgggt gcacacgtgt gcattcatat gcgtgtgtgt gtgtgtgcgt gtgtttatga 11820
gagtgtccat tgctttctcc catggttacc tcctttagaa agaagcagca gtcaggaaga 11880
cagatgtgaa gagctggagc atgttcagat gagaggagac ggaacacggg gacacaccag 11940
cttgagcaag ggacaacagg ggaggactga tgactgactt cccacctttg aggtgctaat 12000
gtgtgtgtgg tggcactgga taaaagatca atgttggcta ggcaccatgg cacacgcctg 12060
tagtcccagc cactctggag gctaaggcgg gaggattgct tgagcccaga agttggaggc 12120
tgctatgagc cgtgatcatg ccactgcact ccagcaacct gggcaacaga gtgagaccct 12180
gtctcaaaaa aaaaaaaaaa aatgaaaagt ccacataacc tgagcatcat gtgcccagag 12240
cgttgggtgg tgtggtccca ttccttcctt ccagcggctt cttctggcca cctcaatgtc 12300
aggatgtcct gctcacatat caataccatt aaaacctgac ttctttccct gcactgttga 12360
agctccttct tgaggctcac attatggata taattttgat tctttcttca gtggtataga 12420
taactacttg taacctaaga acaacttggt gaaagtcctc taatacatta tttttaaaa 12480
aaacacaaat caatgagctc aacttattaa ctaactttca tctattcatt tttgagccat 12540
ccctgtctga ttgtgaatct ccatgattcc aacactctga gctggggata gtgcctacac 12600
aaaataaaaa gaagtggaaa attttcaaac atcagtttat gctgacaacc aggccataat 12660
aggtgctcaa ttactattga atgaatgaat gaaagttctg gccaggtacg gtggctcatg 12720

cctgtagtcc caacactttg ggaggccgag gcaggtggat cacttgaggt taggagttcg 12780
aaaccaacct gaccaacatg aagaaacctt atctctacca aaaaaatata aaaaaattac 12840
ccaggcatgg tggtgtatgc ctgtaatccc agctatttgg gaggctgagg caggaaaatc 12900
acttgaacct gagaggcgga ggttgcagtg agctgagatt gtgccactcc actccagcct 12960
gggcgacaga gtgagactcc gtcttactta aaaaaaaaaa aaagaaggtt ccaagaaaat 13020
tcatcttaag gtttatgtaa aaggaagatg atatttaaca tgattcatgg ccaagtacta 13080
atattacatt ataataatgt ttccaaataa cattatagat atgtttaaag acagtgtatt 13140
aggctgttct tgcattgctg taaagaaata cccaagactg ggtaatttat aaagaaaaga 13200
ggtttcattg gctcgtgttt ctgcaggctg tacaggaagc ttagtgctga catcacttgg 13260
ctgccggggg aacctcaggg agcttttact catggcagaa ggcaatgcgg gagcttgcat 13320
gtcacatggc aaaagcagga gcgagagaga gttgggggg aaggtgccac acacttttta 13380
atgaccggct ctcacaataa ctcatgaaaa ctcactatca ggaagacagc actaaagcac 13440
aagggatccg accccatgat ccaaacacct cccaccaggc cccatctcca gcactgggga 13500
ttacaattca acatgagatc tgagtgtgga caaatatcca aactgtatca gtcaacagcg 13560
atcataatta gtcctgaata ggagtgcctt tttttttctt tcttctccct tttcttttct 13620
acttcctcct ccttttccct ctcctcttca atctcctctt cattcctgta gcaccaaggg 13680
ttgaagcacc taacccgttt tggattgaga tgttctgatt gggcaatgaa cactgtccag 13740
aataaacaga aatccatttt gcactaagtg gctgcacaga ccctgcctca tgctaaatct 13800
agcacccaga tagtttaatg tttcaatgac tgaattacaa atatatcatc accttggatt 13860
tggcacttac aaatggctgt taatttggcc agaggtggtt gtttacaact tcaaatagga 13920
gactattcat aatttctgac gtgacatttt cctttcttta ttttactgta tgaaaatata 13980
atgaaatttc tcacaaaata tcactaaaaa gaaaagaaga agagtaggaa gcaaggttaa 14040
aatatttcta aaatataatt ttggtctttc tttttctccc ttccttcctc cgtccctctc 14100
tcctttcctc tctccctccc tccctccctc ccttcctcct ttccttgctt ccttccctcc 14160
ttctcttcct tctttttcaa gagatcaata acatttatta agaataagtt tcttaattat 14220
aacctttcag gtgataatag taacacagcc tgggcaacac aataagacct tgtttctaca 14280
aaaaatttaa aaattggcca gacatagtgg tgcatgacta attccagcta ctctggaggc 14340
tgaggcagga ggatggcttg agcccaggag ttggaggctg cagttagcca tgcttgtgcc 14400
actacactcc agcccgggca acagggcaag actctgtatc taaaaacaac aacaacaaca 14460
ataatagaaa caggtttcct ttcccaagtt tggaaaatct ggtagtcttc ttaagcagcc 14520
atgagcataa agagaggatt gttcatacca caggtgttcc aggcataacg aaactgtctt 14580
tgtgtttagt tacaaggaga acatcatgcg cctctccagc cttcacaagg accgcccggt 14640
ggagccgctg gacctggccg tgttctgggt ggagtttgtg atgaggcaca agggcgcgcc 14700
acacctgcgc cccgcagccc acgacctcac ctggtaccag taccattcct tggacgtgat 14760
tggtttcctc ttggccgtcg tgctgacagt ggccttcatc acctttaaat gttgtgctta 14820
tggctaccgg aaatgcttgg ggaaaaaagg gcgagttaag aaagcccaca aatccaagac 14880
ccattgagaa gtgggtggga aataaggtaa aattttgaac cattccctag tcatttccaa 14940
acttgaaaac agaatcagtg ttaaattcat tttattctta ttaaggaaat actttgcata 15000
aattaatcag ccccagagtg ctttaaaaaa ttctcttaaa taaaaataat agactcgcta 15060
gtcagtaaag atatttgaat atgtatcgtg ccccctctgg tgtctttgat caggatgaca 15120
tgtgccattt ttcagaggac gtgcagacag gctggcattc tagattactt ttcttactct 15180
gaaacatggc ctgtttggga gtgcgggatt caaaggtggt cccacggctg cccctactgc 15240
aaatggcagt tttaatctta tcttttggct tctgcagatg gttgcaattg atccttaacc 15300
aataatggtc agtcctcatc tctgtcgtgc ttcataggtg ccaccttgtg tgtttaaaga 15360
agggaagctt tgtaccttta gagtgtaggt gaaatgaatg aatggcttgg agtgcactga 15420
gaacagcata tgatttcttg ctttggggaa aaagaatgat gctatgaaat tggtgggtgg 15480
tgtatttgag aagataatca ttgcttatgt caaatggagc tgaatttgat aaaaacccaa 15540
aatacagcta tgaagtgctg ggcaagttta cttttttct gatgtttcct acaactaaaa 15600

ataaattaat aaatttatat aaattctatt taagtgtttt cactggtgtc gcatttattt 15660
cttgttaagt tgcattttct aattacaaaa gtaatgcatg attatgacag aaagtttgga 15720
aaatatagag gttcacacac acacgccttc attgcgtgtg catgcataaa tgcatgagaa 15780
aagaaaaata accagtaatc acatcgccca gaaataaccc cagttacaat tgtggcaaat 15840
acacatactt ataaatattg cagatatatt aagtatacct agtatttgct aacactcttt 15900
cttctactct gtcatgaaga ttctcccaag gtgtttttgt ataatattta attcattttc 15960
agtggccaag cagtattcta cttcatggat ataccaggat ttatttaacc ataacttctg 16020
gttggattca ctcttattat tttgtttaat taaaaaaaaa agacctcggc tgggcacagt 16080
ggctcatgcc tgtaatccca gcactttggg aggccgaggt gggtggatca cctaagatcg 16140
ggagtttgag accagcctgg ccaacatggc aaaaccccgt ctctactaaa aatacagaaa 16200
attagccggg tgtggttgcc agcacctgta attccagcta attgggaggc tgaggcagga 16260
gaattgcttg aacccgggtc aggggttcg gaggtcggag gttgcagtga gtccggatca 16320
tgccactgca ttccagcctg ggtgacacag ccagactctg tctcaaaaac aacaacaaca 16380
acaaaacaac aacaacaaca acaacaaaaa tctcactgga catcctagta gctaaggctt 16440
tccacatatt catgattact tctgttggaa agtgctttac aacaaattgc tagttgtctc 16500
agtctgggtt cccctgagat gaggattcaa gggccaggag tttatttagg aagtaaagga 16560
aacactgata gaggagtggc agagtgagaa ggggtgatgg tcatccacag ctggctctct 16620
tgtggtcaat cggagcttaa tcctgctggg tgactctggg agccagtgga gaaaagacac 16680
cccagactta tcccaatgag gaacacggct gttgggtgcc tgagtacttg cctcgtcagg 16740
gattgaaacg tactcccagg tagtagtaat ttctctgccc ttccattagg ccacaaaggg 16800
ggctctgaca gagagagctg acgagaaaaa acacacgccc ttgtcactga agaggtacac 16860
aggggatctg tgtggggcac cacctgcact gctaccctgg acaaatagct taagaaatcc 16920
ccacactgca tccccaaact tactatcagc gtgtgaggga gacaggttcc cacaccctca 16980
ttagcacaaa gtactatctt gaaaaagaaa gcctgtcagt ttgataggag aaaagcagga 17040
tcttgtttac aatgtgcttt tattattgtt attattagag attgtatttc ttttcaagct 17100
gatgagccgt ctgtgtttat tttttggagg ataccctttg cccactttcc tattggagtg 17160
tattaccctg aggatttggt aagagtgctt attgcattca ccagaatgtc ctttttgtca 17220
tttactgtat tttctctact tttttttttt tgccttgttt tacttttttt gttttgtatt 17280
acaagcagaa gttttaaatt tgtaagcttc aaattggagc tggggtggtg cagagcgaag 17340
atttcagctg gttccctgac cccagctcca tctccttccc taggcagtgg ctggaacaca 17400
ttctgtccac tatttccctc tctacatcct tgaggctgtg cagtcacccc tcaactacgt 17460
tcaccctcct tcaaagccct tcctggtcca cccggggacc atctcccggc ctcactgccc 17520
ctagctcctt gacgccccaa cctctctcag ggaccccaag ttgccatgac ctccagccag 17580
ctcatgttca tttgcacctt cgtgtctgca gcactgaggc actcttgttt acaagtgaga 17640
gaacccaact cgggatacct taagcataaa cagtattttt gtaaggagac aggcttctga 17700
cgacgcgagg ctcatagcca ggcctgcgct gggtggagcc tccccttctc actcctgtcc 17760
ctgttgggtc agagtgccag ctttcctctt ccctctcctc cgggtctttt cggcccctca 17820
gtccccatat tcctctgccc tagctcccaa gatcccacaa gagacagact ggattctctc 17880
tggcctggag tgccaccttc ctgaaagtca gaatctgatt ggtccagctg gatcaggtgt 17940
cctctccctg tccaatcatc aatgccgaga gggattacgg agagaaaaac atggctccca 18000
ccatcccatt gctgtggctg cttgggccac gggagaggga accttgtgag ccgggcagaa 18060
tccaccctgt agagactgcc tctgggtgag tcatatggtt tggctgtgtc cccacccaaa 18120
tctcatcttg aattgtaatt cccattaccc ccatgtgtca ttagagggac ctggtgggaa 18180
gtgatttgaa ttatgggggc agttatctcc atgctatttg tgtgatagtg agttctcaca 18240
agatctgatg gttttatagg gggcttttcc ccctcttgct catacttcct cttgcctgcc 18300
accatgtaag atgtgccctt gctcctcctt caccttctgc catgattgtg aggcctcccc 18360
agccatgtgg aactgtgagt ccatgaaacc tctttttctt tataaattac ccagtcctgg 18420
gtatttcttc atagcagtat gaaaatggac taatacagtc agcttctgca caatatttcc 18480

```
attttcccac attatgtctt gggcctttg tgtatttaag ctcacaggat gctacgaata 18540
aagcgttttc ttatttcctg ggtagttccc atagaagtag tggtgcaacg tgccatagag 18600
tgacagcacc taagagaagc tgattttgtg agtggattgt gagttcaata ttgttgtcat 18660
aatcagaaaa aaatgtattt actttttttt tttttttttt tgagacggag tctcactctg 18720
ttgcccaggc tggagtgcag tggtgtgatc ttggctcact gcaacctccg cctcctgggt 18780
tcaagtgatt cttcctacct caccctcctg agtagctggg attacaggca catgccaccc 18840
caccacaccc ggctaatttt tgtatttttt agtagagaca gcgtttc 18887
```

<210> 2
<211> 1602
<212> DNA
<213> Homo sapiens

<400> 2

```
atggctgtgg agtcccaggg cggacgccca cttgtcctgg gcctgctgct gtgtgtgctg 60
ggcccagtgg tgtcccatgc tgggaagata ctgttgatcc cagtggatgg cagccactgg 120
ctgagcatgc ttggggccat ccagcagctg cagcagaggg gacatgaaat agttgtccta 180
gcacctgacg cctcgttgta catcagagac ggagcatttt acaccttgaa gacgtaccct 240
gtgccattcc aaagggagga tgtgaaagag tcttttgtta gtctcgggca taatgttttt 300
gagaatgatt ctttcctgca gcgtgtgatc aaaacataca agaaaataaa aaaggactct 360
gctatgcttt tgtctggctg ttcccactta ctgcacaaca aggagctcat ggcctccctg 420
gcagaaagca gctttgatgt catgctgacg gaccctttcc ttccttgcag ccccatcgtg 480
gcccagtacc tgtctctgcc cactgtattc ttcttgcatg cactgccatg cagcctggaa 540
tttgaggcta cccagtgccc caacccattc tcctacgtgc ccaggcctct ctcctctcat 600
tcagatcaca tgaccttcct gcagcgggtg aagaacatgc tcattgcctt ttcacagaac 660
tttctgtgcg acgtggttta ttccccgtat gcaacccttg cctcagaatt ccttcagaga 720
gaggtgactg tccaggacct attgagctct gcatctgtct ggctgtttag aagtgacttt 780
gtgaaggatt accctaggcc catcatgccc aatatggttt ttgttggtgg aatcaactgc 840
cttcaccaaa atccactatc ccaggaattt gaagcctaca ttaatgcttc tggagaacat 900
ggaattgtgg ttttctcttt gggatcaatg gtctcagaaa ttccagagaa gaaagctatg 960
gcaattgctg atgctttggg caaaatccct cagacagtcc tgtggcggta cactggaacc 1020
cgaccatcga atcttgcgaa caacacgata cttgttaagt ggctacccca aaacgatctg 1080
cttggtcacc cgatgacccg tgcctttatc acccatgctg gttcccatgg tgtttatgaa 1140
agcatatgca atggcgttcc catggtgatg atgcccttgt ttggtgatca gatggacaat 1200
gcaaagcgca tggagactaa gggagctgga gtgaccctga atgttctgga aatgacttct 1260
gaagatttag aaaatgctct aaaagcagtc atcaatgaca aaagttacaa ggagaacatc 1320
atgcgcctct ccagccttca caaggaccgc ccggtggagc cgctggacct ggccgtgttc 1380
tgggtggagt ttgtgatgag gcacaagggc gcgccacacc tgcgccccgc agcccacgac 1440
ctcacctggt accagtacca ttccttggac gtgattggtt tcctcttggc cgtcgtgctg 1500
acagtggcct tcatcacctt taaatgttgt gcttatggct accggaaatg cttggggaaa 1560
aaagggcgag ttaagaaagc ccacaaatcc aagacccatt ga 1602
```

<210> 3
<211> 533
<212> PRT
<213> Homo sapiens

```
<400> 3
Met Ala Val Glu Ser Gln Gly Gly Arg Pro Leu Val Leu Gly Leu Leu
 1               5                  10                  15

Leu Cys Val Leu Gly Pro Val Val Ser His Ala Gly Lys Ile Leu Leu
            20                  25                  30

Ile Pro Val Asp Gly Ser His Trp Leu Ser Met Leu Gly Ala Ile Gln
        35                  40                  45

Gln Leu Gln Gln Arg Gly His Glu Ile Val Val Leu Ala Pro Asp Ala
    50                  55                  60

Ser Leu Tyr Ile Arg Asp Gly Ala Phe Tyr Thr Leu Lys Thr Tyr Pro
65                  70                  75                  80

Val Pro Phe Gln Arg Glu Asp Val Lys Glu Ser Phe Val Ser Leu Gly
                85                  90                  95

His Asn Val Phe Glu Asn Asp Ser Phe Leu Gln Arg Val Ile Lys Thr
            100                 105                 110

Tyr Lys Lys Ile Lys Lys Asp Ser Ala Met Leu Leu Ser Gly Cys Ser
        115                 120                 125

His Leu Leu His Asn Lys Glu Leu Met Ala Ser Leu Ala Glu Ser Ser
    130                 135                 140

Phe Asp Val Met Leu Thr Asp Pro Phe Leu Pro Cys Ser Pro Ile Val
145                 150                 155                 160

Ala Gln Tyr Leu Ser Leu Pro Thr Val Phe Phe Leu His Ala Leu Pro
                165                 170                 175

Cys Ser Leu Glu Phe Glu Ala Thr Gln Cys Pro Asn Pro Phe Ser Tyr
            180                 185                 190

Val Pro Arg Pro Leu Ser Ser His Ser Asp His Met Thr Phe Leu Gln
        195                 200                 205

Arg Val Lys Asn Met Leu Ile Ala Phe Ser Gln Asn Phe Leu Cys Asp
    210                 215                 220

Val Val Tyr Ser Pro Tyr Ala Thr Leu Ala Ser Glu Phe Leu Gln Arg
225                 230                 235                 240

Glu Val Thr Val Gln Asp Leu Leu Ser Ser Ala Ser Val Trp Leu Phe
```

245 250 255

Arg Ser Asp Phe Val Lys Asp Tyr Pro Arg Pro Ile Met Pro Asn Met
260 265 270

Val Phe Val Gly Gly Ile Asn Cys Leu His Gln Asn Pro Leu Ser Gln
275 280 285

Glu Phe Glu Ala Tyr Ile Asn Ala Ser Gly Glu His Gly Ile Val Val
290 295 300

Phe Ser Leu Gly Ser Met Val Ser Glu Ile Pro Glu Lys Lys Ala Met
305 310 315 320

Ala Ile Ala Asp Ala Leu Gly Lys Ile Pro Gln Thr Val Leu Trp Arg
325 330 335

Tyr Thr Gly Thr Arg Pro Ser Asn Leu Ala Asn Asn Thr Ile Leu Val
340 345 350

Lys Trp Leu Pro Gln Asn Asp Leu Leu Gly His Pro Met Thr Arg Ala
355 360 365

Phe Ile Thr His Ala Gly Ser His Gly Val Tyr Glu Ser Ile Cys Asn
370 375 380

Gly Val Pro Met Val Met Met Pro Leu Phe Gly Asp Gln Met Asp Asn
385 390 395 400

Ala Lys Arg Met Glu Thr Lys Gly Ala Gly Val Thr Leu Asn Val Leu
405 410 415

Glu Met Thr Ser Glu Asp Leu Glu Asn Ala Leu Lys Ala Val Ile Asn
420 425 430

Asp Lys Ser Tyr Lys Glu Asn Ile Met Arg Leu Ser Ser Leu His Lys
435 440 445

Asp Arg Pro Val Glu Pro Leu Asp Leu Ala Val Phe Trp Val Glu Phe
450 455 460

Val Met Arg His Lys Gly Ala Pro His Leu Arg Pro Ala Ala His Asp
465 470 475 480

Leu Thr Trp Tyr Gln Tyr His Ser Leu Asp Val Ile Gly Phe Leu Leu
485 490 495

Ala Val Val Leu Thr Val Ala Phe Ile Thr Phe Lys Cys Cys Ala Tyr

500 505 510

Gly Tyr Arg Lys Cys Leu Gly Lys Lys Gly Arg Val Lys Lys Ala His
515 520 525

Lys Ser Lys Thr His
530

<210> 4
<211> 15
<212> DNA
<213> Homo sapiens

<400> 4
cttttayag tcacg 15

<210> 5
<211> 15
<212> DNA
<213> Homo sapiens

<400> 5
gggccatyca gcagc 15

<210> 6
<211> 15
<212> DNA
<213> Homo sapiens

<400> 6
gcctggartt tgagg 15

<210> 7
<211> 15
<212> DNA
<213> Homo sapiens

<400> 7
tgctgagyaa gcatt 15

<210> 8
<211> 15
<212> DNA

<213> Homo sapiens

<400> 8
gattctayac catgg 15

<210> 9
<211> 15
<212> DNA
<213> Homo sapiens

<400> 9
tctatacyat ggcct 15

<210> 10
<211> 15
<212> DNA
<213> Homo sapiens

<400> 10
cagaggaycc ctgtt 15

<210> 11
<211> 15
<212> DNA
<213> Homo sapiens

<400> 11
aggacccytg ttttc 15

<210> 12
<211> 15
<212> DNA
<213> Homo sapiens

<400> 12
ggaccccygt tttct 15

<210> 13
<211> 15
<212> DNA
<213> Homo sapiens

<400> 13

tattatgytc tttct 15

<210> 14
<211> 15
<212> DNA
<213> Homo sapiens

<400> 14
gggcaacwgg gcaag 15

<210> 15
<211> 15
<212> DNA
<213> Homo sapiens

<400> 15
tgcgcccygc agccc 15

<210> 16
<211> 15
<212> DNA
<213> Homo sapiens

<400> 16
tcttggcygt cgtgc 15

<210> 17
<211> 15
<212> DNA
<213> Homo sapiens

<400> 17
aattcatytt attct 15

<210> 18
<211> 15
<212> DNA
<213> Homo sapiens

<400> 18
tgacagcttt ttaya 15

<210> 19
<211> 15
<212> DNA
<213> Homo sapiens

<400> 19
gtgtcacgtg actrt 15

<210> 20
<211> 15
<212> DNA
<213> Homo sapiens

<400> 20
tgcttggggc catyc 15

<210> 21
<211> 15
<212> DNA
<213> Homo sapiens

<400> 21
gctgcagctg ctgra 15

<210> 22
<211> 15
<212> DNA
<213> Homo sapiens

<400> 22
catgcagcct ggart 15

<210> 23
<211> 15
<212> DNA
<213> Homo sapiens

<400> 23
gggtagcctc aaayt 15

<210> 24
<211> 15
<212> DNA

<213> Homo sapiens

<400> 24
gatatatgct gagya 15

<210> 25
<211> 15
<212> DNA
<213> Homo sapiens

<400> 25
tctcagaatg cttrc 15

<210> 26
<211> 15
<212> DNA
<213> Homo sapiens

<400> 26
taagaagatt ctaya 15

<210> 27
<211> 15
<212> DNA
<213> Homo sapiens

<400> 27
atgaggccat ggtrt 15

<210> 28
<211> 15
<212> DNA
<213> Homo sapiens

<400> 28
gaagattcta tacya 15

<210> 29
<211> 15
<212> DNA
<213> Homo sapiens

<400> 29

gatatgaggc catrg 15

<210> 30
<211> 15
<212> DNA
<213> Homo sapiens

<400> 30
ttccttcaga ggayc 15

<210> 31
<211> 15
<212> DNA
<213> Homo sapiens

<400> 31
ctagaaaaca gggrt 15

<210> 32
<211> 15
<212> DNA
<213> Homo sapiens

<400> 32
cttcagagga cccyt 15

<210> 33
<211> 15
<212> DNA
<213> Homo sapiens

<400> 33
taactagaaa acarg 15

<210> 34
<211> 15
<212> DNA
<213> Homo sapiens

<400> 34
ttcagaggac cccyg 15

<210> 35
<211> 15
<212> DNA
<213> Homo sapiens

<400> 35
ctaactagaa aacrg 15

<210> 36
<211> 15
<212> DNA
<213> Homo sapiens

<400> 36
taatcatatt atgyt 15

<210> 37
<211> 15
<212> DNA
<213> Homo sapiens

<400> 37
acgtaaagaa agarc 15

<210> 38
<211> 15
<212> DNA
<213> Homo sapiens

<400> 38
cagcccgggc aacwg 15

<210> 39
<211> 15
<212> DNA
<213> Homo sapiens

<400> 39
cagagtcttg cccwg 15

<210> 40
<211> 15
<212> DNA

<213> Homo sapiens

<400> 40
cacacctgcg cccyg 15

<210> 41
<211> 15
<212> DNA
<213> Homo sapiens

<400> 41
ggtcgtgggc tgcrg 15

<210> 42
<211> 15
<212> DNA
<213> Homo sapiens

<400> 42
gtttcctctt ggcyg 15

<210> 43
<211> 15
<212> DNA
<213> Homo sapiens

<400> 43
ctgtcagcac gacrg 15

<210> 44
<211> 15
<212> DNA
<213> Homo sapiens

<400> 44
gtgttaaatt catyt 15

<210> 45
<211> 15
<212> DNA
<213> Homo sapiens

<400> 45

ttaataagaa taara 15

<210> 46
<211> 10
<212> DNA
<213> Homo sapiens

<400> 46
cagcttttta 10

<210> 47
<211> 10
<212> DNA
<213> Homo sapiens

<400> 47
tcacgtgact 10

<210> 48
<211> 10
<212> DNA
<213> Homo sapiens

<400> 48
ttggggccat 10

<210> 49
<211> 10
<212> DNA
<213> Homo sapiens

<400> 49
gcagctgctg 10

<210> 50
<211> 10
<212> DNA
<213> Homo sapiens

<400> 50
gcagcctgga 10

<210> 51
<211> 10
<212> DNA
<213> Homo sapiens

<400> 51
tagcctcaaa 10

<210> 52
<211> 10
<212> DNA
<213> Homo sapiens

<400> 52
atatgctgag 10

<210> 53
<211> 10
<212> DNA
<213> Homo sapiens

<400> 53
cagaatgctt 10

<210> 54
<211> 10
<212> DNA
<213> Homo sapiens

<400> 54
gaagattcta 10

<210> 55
<211> 10
<212> DNA
<213> Homo sapiens

<400> 55
aggccatggt 10

<210> 56
<211> 10
<212> DNA

<213> Homo sapiens

<400> 56
gattctatac 10

<210> 57
<211> 10
<212> DNA
<213> Homo sapiens

<400> 57
atgaggccat 10

<210> 58
<211> 10
<212> DNA
<213> Homo sapiens

<400> 58
cttcagagga 10

<210> 59
<211> 10
<212> DNA
<213> Homo sapiens

<400> 59
gaaaacaggg 10

<210> 60
<211> 10
<212> DNA
<213> Homo sapiens

<400> 60
cagaggaccc 10

<210> 61
<211> 10
<212> DNA
<213> Homo sapiens

<400> 61

ctagaaaaca 10

<210> 62
<211> 10
<212> DNA
<213> Homo sapiens

<400> 62
agaggacccc 10

<210> 63
<211> 10
<212> DNA
<213> Homo sapiens

<400> 63
actagaaaac 10

<210> 64
<211> 10
<212> DNA
<213> Homo sapiens

<400> 64
tcatattatg 10

<210> 65
<211> 10
<212> DNA
<213> Homo sapiens

<400> 65
taaagaaaga 10

<210> 66
<211> 10
<212> DNA
<213> Homo sapiens

<400> 66
cccgggcaac 10

<210> 67
<211> 10
<212> DNA
<213> Homo sapiens

<400> 67
agtcttgccc 10

<210> 68
<211> 10
<212> DNA
<213> Homo sapiens

<400> 68
acctgcgccc 10

<210> 69
<211> 10
<212> DNA
<213> Homo sapiens

<400> 69
cgtgggctgc 10

<210> 70
<211> 10
<212> DNA
<213> Homo sapiens

<400> 70
tcctcttggc 10

<210> 71
<211> 10
<212> DNA
<213> Homo sapiens

<400> 71
tcagcacgac 10

<210> 72
<211> 10
<212> DNA

<213> Homo sapiens

<400> 72
ttaaattcat 10

<210> 73
<211> 10
<212> DNA
<213> Homo sapiens

<400> 73
ataagaataa 10

<210> 74
<211> 18887
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (2510)
<223> PS1: polymorphic base T or C

<220>
<221> allele
<222> (2756)
<223> PS2: polymorphic base C or T

<220>
<221> allele
<222> (2826)
<223> PS3: polymorphic base G or A

<220>
<221> allele
<222> (3155)
<223> PS4: polymorphic base A or G

<220>
<221> allele
<222> (3568)
<223> PS5: polymorphic base C or T

<220>
<221> allele
<222> (9508)

<223> PS6: polymorphic base T or C

<220>
<221> allele
<222> (9511)
<223> PS7: polymorphic base C or T

<220>
<221> allele
<222> (10091)
<223> PS8: polymorphic base C or T

<220>
<221> allele
<222> (10094)
<223> PS9: polymorphic base C or T

<220>
<221> allele
<222> (10095)
<223> PS10: polymorphic base T or C

<220>
<221> allele
<222> (10140)
<223> PS11: polymorphic base T or C

<220>
<221> allele
<222> (14423)
<223> PS12: polymorphic base A or T

<220>
<221> allele
<222> (14713)
<223> PS13: polymorphic base C or T

<220>
<221> allele
<222> (14776)
<223> PS14: polymorphic base C or T

<220>
<221> allele
<222> (14971)
<223> PS15: polymorphic base T or C

<400> 74

gaattcaagg gattcaagga aggtggcttt gcttcccggg agggtcctgt agatgatcta 60
cagggcactg gacatgttta tgttgctcct ttagtaataa gcctgtcatt ctgatttgat 120
gaaaggagat gaaaggagct ggtagtgtgt ctgatggtgg cctactaact tatgtcttca 180
gcttaaaaag aaagtagctt caaaagggtt ccagaaacac tttccatgga cgtgtcactc 240
tttagcagcc cccaaagcaa gaccatcata ttgctgccct gctgtgtgat ttctcagccc 300
ctagagcacc atcccctgta attgcctggt catgagtttg tctctgtcta cctgacccct 360
cctttcaggc aaggaccatt tctaacttga ctttctgggc ctagttccta gcatagtgac 420
tgccatccag tagggctcac acgttccata aatatttggc agatgaggga attagcaatg 480
ggttctgctt tggtttcaga gcagatatta attggattgc ttagtagtgg ttctctgttg 540
taattcatga gcatgaatgt ggattgccca ctattcagat tagtaagtat ttcttggtca 600
agggcagagc tgtggccaca aaccatccag gtacacagca gaagcagcct caaaaagctt 660
ggaagctctg catgatgcag gaaagtcata aaatcattac agtggtgact tatgtgttta 720
tagccccttt actgtctata atctgcaaat gaactcacac agcattggga ctttggaaga 780
attatcaccc ttaaggttta aattaaactg tgaatttcag aatttctaat aaggacacaa 840
caaagagtga aagcattgct atgtctattc tgcttgccca gaatcttggt cctaaaaaat 900
gaagagtgtt tgggtgtggg gaggagcttc agtgtgcatg tgcatgcaaa gtacctactc 960
taaggagaag aatgagaggg taccctaatt acctgttaat atgtcccata ggacaccaaa 1020
actctagtta gctgtttctc tatgatcctc taagcacatc cccaagtatg gctggccagt 1080
gatgtgtatg gttcaaatgt tgggatctgt gcagttatct tggaattgta tagtacagca 1140
gtatatcccc cccaaaaaga gtgtaatact tccaattctg gctgcacaat acttgcccca 1200
tagtccatgg tcaataaata caaatttgag ttgtttttgc tcatctttcc cttttgactt 1260
caaatcagtc atcagaattt ccccaaatgc ctttcccctg gatcttgggc cagtggaatg 1320
agtacaattt aacttaattg aatttgctta tctatttggt ttcctgttgt gaacaaaagt 1380
tctctgaaaa ggaatttgga agaaagagac tttgttctag tgaacagttt gcaaaccagg 1440
gagttacagc ctctggtacg caatgaaggt gagttccaca gaacacaagg caggcaggtt 1500
tcacggcaaa aagttccttc ccaggttccc aatcaggtcc atttatgcaa atgaaggatg 1560
gaaacttgct tagttcttat tggtcactgc agctgcattc tgattggttg atgaagctga 1620
gccctgagtg gctgaggtgg gtgagcttta attggttggt tcaggtgagc gctgaaaatc 1680
tcaactataa aaaggtacag gttttcagga tactcagagt aaccgtgtga cctgtagtaa 1740
gcaaagggcc agttggctct attttaaatc caggcccagt tagccactca agatctatct 1800
tacaggactg gctctttcag gttcacacta ataaaggcct gtccttgggg aagacttctg 1860
ttcacatgcg ctccagtgaa tttccctttc tggtcattct ctaccccagc acgcccccca 1920
cccccgaccc gccccaccca cccacctgtt catttccttc ttagcatgct tcacgatttc 1980
taagttcctg ctcatgtgtt taaattgtga gtctggctca cctcatggcg cgtgctcgtg 2040
tggtgggctc tgctgcagcc tcaagacccc acactgtgct ggactcaata aatattgttg 2100
gacgaaggaa tgaaacacat gatacaagtg agcaggcagt accgggggag ctgtggagtg 2160
ggcactctta caggtttcca tggcgaaagc gggggtacag ttgtgttctt ttctttctaa 2220
aaggctttct aaaaagcctt ctgtttaatt tctggaaaag aagcctaact tgttcactac 2280
atagtcgtcc ttcttcctct ctggtaacac ttgttggtct gtggaaatac taatttaatg 2340
gatcctgagg ttctggaagt actttgctgt gttcactcaa gaatgtgatt tgagtatgaa 2400
attccagcca gttcaactgt tgttgcctat taagaaacct aataaagctc caccttcttt 2460
atctctgaaa gtgaactccc tgctaccttt gtggactgac agctttttay agtcacgtga 2520
cacagtcaaa cattaacttg gtgtatcgat tggtttttgc catatatata tatataagta 2580
ggagagggcg aacctctggc aggagcaaag gcgccatggc tgtggagtcc cagggcggac 2640
gcccacttgt cctgggcctg ctgctgtgtg tgctgggccc agtggtgtcc catgctggga 2700
agatactgtt gatcccagtg gatggcagcc actggctgag catgcttggg gccatycagc 2760
agctgcagca gaggggacat gaaatagttg tcctagcacc tgacgcctcg ttgtacatca 2820
gagacrgagc attttacacc ttgaagacgt accctgtgcc attccaaagg gaggatgtga 2880

aagagtcttt tgttagtctc gggcataatg tttttgagaa tgattctttc ctgcagcgtg 2940
tgatcaaaac atacaagaaa ataaaaaagg actctgctat gcttttgtct ggctgttccc 3000
acttactgca caacaaggag ctcatggcct ccctggcaga aagcagcttt gatgtcatgc 3060
tgacggaccc tttccttcct tgcagcccca tcgtggccca gtacctgtct ctgcccactg 3120
tattcttctt gcatgcactg ccatgcagcc tggartttga ggctacccag tgccccaacc 3180
cattctccta cgtgcccagg cctctctcct ctcattcaga tcacatgacc ttcctgcagc 3240
gggtgaagaa catgctcatt gccttttcac agaactttct gtgcgacgtg gtttattccc 3300
cgtatgcaac ccttgcctca gaattccttc agagagaggt gactgtccag gacctattga 3360
gctctgcatc tgtctggctg tttagaagtg actttgtgaa ggattaccct aggcccatca 3420
tgcccaatat ggtttttgtt ggtggaatca actgccttca ccaaaatcca ctatcccagg 3480
tgtgtattgg agtgggactt ttacatgcgt atattctttc agatgtatta ctttggatcg 3540
attaactagc cccagatata tgctgagyaa gcattctgag ataatttaaa atgccctctt 3600
ttgttaattt ttgactccta ggtttgagtc tgtctttggc atcatcttct ggatgatttc 3660
ttggtatctg agatttcggg aaagcattcc ttggacattt tactctgtgt gctccagtgg 3720
atagtaatca attagaaaca acaagctgtt aaatgccata ggcacagaat gctgggtttg 3780
gggcaccctg cagaaaactc agttgaagcc tgcaccttgc cctggattca gtcaggcagg 3840
caatgttcag gactgatgaa atcattcttt gatgatgata gatcctggaa atgaaagttg 3900
cctttgtgac cctggttaaa gctccagttt ctaaatattc tgataagaag ctaaatcctg 3960
cagtccgttc tcttctaatg agtgaatcac cagacagtca ggttctgaca tgatacagaa 4020
aggttgtagg tttcattctc aagctattag gtttattttt cccctacaga gtttgaagta 4080
tgcaaaaagt agcattcaca tcctcatcga aatctcagca gaggatagaa aagaacagga 4140
gaggctcctt cagatggagc gttagggaat tactctttga ggaggtgaca tttcagagag 4200
cgttcattca cttatcctgc aaagattggc tgaggatcta ctggcagccc aggcacttcc 4260
caggtgctgc gtctggctcc cattaagggg actgatatca ccttcggagg tgaccttatt 4320
tccactatac ctccaatgtg atttgtattt tatttttttt aattttctgt gcattttcct 4380
tcatagcaca tcaaatatgg cagccatttc acttagatag ttgttgattg tccgcttcac 4440
atcatgagcc atgtggggac ctgtgtgact ttgcattaat cacatccact gtatgcggcg 4500
tcctcaacac ctgccaatgg gtctgcatgt atttggcgcc ccataaatct cagcacctaa 4560
ggcacagaat aggcacccac cgaatatgtg ttacattaat gaatgagaag aaaggtgcca 4620
accgaggtct agttaatggg tcgagagtaa tccacaatag ctctttttag ttctttgtac 4680
tccagctatt acataccaat atgtatatag aaacatatgt aaaatttttt ggttgctttt 4740
tctacaaaat agagtaacag tgtattccca ctgcccactt accgataatg tcatggatat 4800
cactccagtt ttaaatgcta ttacttttta aactatgaaa tagtatttca tggtacttgt 4860
gtaccacagt gtattctgct ggagatctag tctagttccc cacagaggaa cattacaatt 4920
tgtattccag gagttttgtt gttgtgacct caaacacttc ctttaaaaag ataagctatt 4980
ttgtagttta aaaaacattt gttctgtttc tttctcattc atcttttctt aagtatttta 5040
cacggttttt tttttttggt cactactgtg aatgtgttat ttttttgcat ttctatctct 5100
agctgattat ctactcatta ctcagctatc tcatcaaaat attgattttc ataataaaaa 5160
ataataggca gtcatttgct gataaagaaa ttttggtttc ttctcttata aattccatgc 5220
caaatatcag ggctattgaa tttattagaa tctctaaaaa cagttgaata attctggcaa 5280
taggaaagat gcccgtcttg ctgctatttt agtggaaatt gattatcatt tcattatttt 5340
gcattatgtt agccattgtt ttctgaacag gctttattga tttagataat ttccttcttt 5400
gcgtgaggat gtttgtagga gaggcaccga actttatcag ctgcctttct ggcatttatt 5460
gatataacca taaaagtcta agtggtgaac tgtgttgact acatatttgt tgttgccttg 5520
tttggtgcag tcaggcttag gtgtgaaaat atgtttttaa attgtacctt ttagtaacct 5580
gttttgtctt gttgcatgtt ttaatctgaa attccacttt ttggatatta atattaccac 5640
ttctgtatta tttttgttta catttcccta gcacatcttt agtactcctt tgtcttcaag 5700
ctttcttcct ttttaaacaa catggcactg gtatttttaa tccagtcagg cagttgcttt 5760

aataagtgca ttttgcctat ttgaatctaa caattaatag atttgattgt aactctctca 5820
gtttacttta tgtttagttg actttgccat tctccttttt ccggatttct actggttggt 5880
caagttactg ttcttatttt ctctttcttc ctttgttaac taaaaatgcc actctgcact 5940
accattcctc ttgtgttgat ggtcctattc tcaatactct tgataaaact cctgaacttt 6000
aagaataaag ataaaacttt tattgcacaa agaagtccat agagaaagca caacctggca 6060
ttggcgtgtc tttggtgtgt ctgaaggaaa agagatagtg gaacaacatt gggagaaaag 6120
gaatgaaact caagaattcc aagatgttcc tcccctgcca gggtaagata gcagtggttc 6180
acagacaatc gcaatgctgg gtctgagaaa aataactaaa cagaagatta gtgaggacca 6240
aggcttcgag atggccagga gaggaaagct tgggagcagg gaaggttgag atatatgtgg 6300
gttactggga atgcgtgatg gtgaagtcac agatgaccca catggtgtct aagtgctaaa 6360
gaagaattct gggaaaatga aatgcatttg ggaagggaaa atctaattaa aagcctaaac 6420
taaaaataca aaattcttgg taaagtttag gagttatgtt aaatgtctca ttttggctgg 6480
tgaagtctca tcagaacagg gaaattctct cattcagggg catctcatct tttctttgaa 6540
gggaatcaat ggtgggggat tggagtgtta ttttcagtta atatgttgct tcactctttg 6600
gtcattccgg taactgtgaa gtcagggtga agtttaaggg aagctttgcc aagtagggga 6660
tggacttcac ctttattgag cctcatagta gctggctcag gtaggagttg gccgtgatga 6720
caacttctct gcagtttgcc ctgcgtgaat ctccagatga acttttgtgc catttaaact 6780
ttcgtgatct cctgctattt aacttcgaat gtttatggac ctgtgggttc aattttgtgt 6840
gaatcacatc ctgctgattg ctgagtgggc gtgtgggagg gtgtgcctgg aggagaactt 6900
agactcggcc ttttccagat gagcttcagt gtaagagtgg gtttcatgaa gagcaaaggt 6960
cctaggaaat ttaagtaagc catttaccaa cgctcagaag aaagaacttg aagagcactt 7020
ggaaatgagc tgtgtctccc caagaaagag ggagagaaag aggggagaga tgtggtgcag 7080
accctaggga ggaaggagtt cagaaaaacc atcctcaggg tgttcttgct acaaaccaaa 7140
aaatgcagca tggtggtggg gaggatgact ctgtcctccc tgacttttag atgagcccaa 7200
gggaaaaggc aaagacaaag cccttaagag ccagaggact cacgagggcc tggggctggt 7260
gagagtggcg gggagagagg gctcaccttg ggagaaggat ggtcagtgtc tggggctttc 7320
ctggtcatgt tccaaatcag gcttggcagg agtcctgctg tgcaaattgc gtttgctgag 7380
ccctgtcaga ggtctcctgt gtctcacatc tagggtgacc agcatcctgg cttcctcagg 7440
actgttcagg ttttagcact gaacatcaca tgtcctaggg aacccctcag tttgggcaag 7500
ccctgccaca tcacacaatc atattagtgc cctcagtatt ctttgcaaac ataaaaccat 7560
agactcagta atcccattac tgggtatata cccaaagaaa tataaattat tctactataa 7620
agacacatgc acatatttgt ttattgcagc actattcaca ataacaaagt catggaacca 7680
acccagatgc ccatcaatgg tagattggat aaagaaaatg tggtacatat acaccatgga 7740
atactatgca gccataacaa ggaatgagat catattcttt gcaaggacat ggatgaagct 7800
ggaagccatc atcctccaca aactaacaca ggaacagaaa atcaaacacc gcatgttctc 7860
actcataagt gggagttgaa cagtgagaat gcgtagacgc agggaggggga acaacacaca 7920
ccagggcttg tggcggggtg aggggtgagg ggaggaactt agaggatagg tcaataggtg 7980
cagcaaacca ccatggcata tgtatcccag aacttcaagt aaataataat aataataatt 8040
aataataata ataataataa ataaacccat aaagccattt gagagattct tgggggattc 8100
attggaccac tgaaaatcta cagtgagaaa agaattgcca tgttgatgaa acaggaaaac 8160
tttccttgtc ccccctcacag agcatgtgac agcgggaggg gctcactttc tcagtgcgcc 8220
actgctcaaa cctctagggg agcatacaga cgggcaggtt gtggggctct gacctcaccg 8280
gcagtgtcta gaggtggatg tttacagctc ctgaagctcc agtgggcgtg ggttatggcc 8340
ttcttttagt tttgccctct atagtcagct tgtgttaacc agctcaatta caccctctac 8400
cttgtcgcaa ggacagaggg ctttctgtat cctgggggct tgccttggtg taccagaaga 8460
atcgaatccc acctgggctt ggagaatgag tgcaaggatt tattgagtgg atgtagctct 8520
cagcagatgg gggaagccag aaggggatgg aatgggaagg gtttcccctg gagtcagacc 8580
gctcagtggc ccgggctcgg tggcccgggc tcggtggcct gggctctcct ccgactgcct 8640

cagccaaact ccgcgttgtt ctgctggtca gtggcctgcc ggtgcctgtt ggtgagttct 8700
tctcaatgtc cagctgtcct tgcgtccctc cgctgatgtg ctcctcccga tgtccagcta 8760
cctgtgtgtc tgcctgctag ggtcttgggg tttttatagg cacatgatgg gggcgtggca 8820
ggccagggtg gttttgggaa atgaaacatt taggcaggaa aacaaaaatg cctgtcctca 8880
cctaggtcca tgggcacagg tctgggggtg gagccctcgc cagggaccac accctcttct 8940
acccagcact tcccttccct acttccatat catttaaagg gaccacgccc ttcccagctc 9000
ttcccttctg tatcactgat gccttgctct gtgttctcta agtggaatta tcactgtgtg 9060
tatgtacagg tgtgtgcatg tgtgtgcatg tacctgtgct tttcttttgg aaaactagca 9120
cattacctgg attttgcatc tcaaggataa ttctgtaagc aggaaccctt cctcctttag 9180
aaggaagtaa aggagaggaa aatgctgtaa aacttacata ttaataattt tttactctat 9240
ctcaaacacg catgccttta atcatagtct taagaggaag atatctaatt cataacttac 9300
tgtatgtagt catcaaagaa tatgagaaaa aattaactga aaatttttct tctggctcta 9360
ggaatttgaa gcctacatta atgcttctgg agaacatgga attgtggttt tctctttggg 9420
atcaatggtc tcagaaattc cagagaagaa agctatggca attgctgatg ctttgggcaa 9480
aatccctcag acagtaagaa gattctayac yatggcctca tatctatttt cacaggagcg 9540
ctaatcccag acttccagct tccagattaa ttctcttaat tggaacctta gatttggctt 9600
ttccctgcca cttcccaact attaatccaa aggttttttt tgttgttgtg gttgttgtca 9660
ttgttttcaa tttgactctc aaatactcta ttaaactatg atccaccaca ctcagaagta 9720
tcattttctc taagagactc aaaagtgtat tagggagaat ttatttaaaa ataaaataaa 9780
tgggatattg tttcttcata ttaaatagaa gtatttctcc aaaaagctgt tggttagaac 9840
actgaattta tgtcttacat ttctgctctt atagttctgc atccacttgt ttcattaagc 9900
aaactttccc ttaaagtgca ggaaagtgaa aaaatcctaa gtgcacagct tgataaatta 9960
tcacaaattc acgtagtgca tacacccttg taactaaacc tccaaaacaa gatgccggaa 10020
gttgccagtc ctcagaagcc ttcacagtta ctgatcctcc cactctgtta aagactgttc 10080
cttcagagga yccyygtttt ctagttagta tagcagattt gttttctaat catattatgy 10140
tctttcttta cgttctgctc tttttgcccc tcccaggtcc tgtggcggta cactggaacc 10200
cgaccatcga atcttgcgaa caacacgata cttgttaagt ggctacccca aaacgatctg 10260
cttggtatgt tgggcggatt ggatgtatag gtcaaaccag ggtcaaatta agaaaatggc 10320
ttaagcacag ctattctaaa ggattgttga gcttgaaaat attatggcca acatatccta 10380
cattgctttt tatctagtgg ggtatctcaa cccacatttt cttctgcaaa tttctgcaag 10440
ggcatgtgag taacactgag tctttggagt gttttcagaa cctagatgtg tccagctgtg 10500
aaactcagag atgtaactgc tgacatcctc cctattttgc atctcaggtc acccgatgac 10560
ccgtgccttt atcacccatg ctggttccca tggtgtttat gaaagcatat gcaatggcgt 10620
tcccatggtg atgatgccct tgtttggtga tcagatggac aatgcaaagc gcatggagac 10680
taagggagct ggagtgaccc tgaatgttct ggaaatgact tctgaagatt tagaaaatgc 10740
tctaaaagca gtcatcaatg acaaaaggta agaaagaaga tacagaagaa tactttggtc 10800
atggcattca tgataaaatt gtttcaaata tgaaaacatt tacgtagcat ttaatagcgt 10860
tgtttcaaat ataaaaacaa atacataaaa atctggattt ttatttcttc tttttttttt 10920
tttttttttt ttgagatgga gtcttgctct gtcacctagg ctggagtgca gtggtgcaat 10980
cttggcttac tgcaacctcc acctcccacg ttcaagcagt tctgcctcag cctccgtgta 11040
gctgggatta caggtgtcca ccaccacgcc cggttaattt ttgtattttt tagtagagaa 11100
agggtttcac catgtttgtc aggctggtct tgaactcctg acttcaggtg atccacctgc 11160
ctcggcctgc caaagtgctg agattacagg catgagccag cgcgtctgac ctggatttat 11220
aaataagata atttagaggt tattattcac tttataaaag gattctttag tttctatata 11280
atttatcata taatttattt agaattttat ttcccccatt agatttaaaa ctccaattta 11340
cataaaaagt tgccataata gacatctgat ccataagttt cctgcacaga aagaaatact 11400
ccattataag aagcatagta tctttaagag aaaaacaact caaatgctta gaagtacagc 11460
tttttgcagc actggaacct gtgagaaatt ttgtccatgg agtttatgaa tgaaggagct 11520

ataagatatc acagacaaag tcttagaata agagcaaagg aaaatttgct caaatgtggc 11580
cctgaaaacg attcaaaggg caaatgattt ctggattaaa gttagtatat tactgtcaag 11640
ctcactggta ataggcttat tagaacctta tgggaagaag tggtggccag tggtagattt 11700
catccgacaa tagatactgt gtgcatatgt gcgtgtgcgt ttgtgcatgt ggctgtgctc 11760
atgtgtgggt gcacacgtgt gcattcatat gcgtgtgtgt gtgtgtgcgt gtgtttatga 11820
gagtgtccat tgctttctcc catggttacc tcctttagaa agaagcagca gtcaggaaga 11880
cagatgtgaa gagctggagc atgttcagat gagaggagac ggaacacggg gacacaccag 11940
cttgagcaag ggacaacagg ggaggactga tgactgactt cccacctttg aggtgctaat 12000
gtgtgtgtgg tggcactgga taaaagatca atgttggcta ggcaccatgg cacacgcctg 12060
tagtcccagc cactctggag gctaaggcgg gaggattgct tgagcccaga agttggaggc 12120
tgctatgagc cgtgatcatg ccactgcact ccagcaacct gggcaacaga gtgagaccct 12180
gtctcaaaaa aaaaaaaaaa aatgaaaagt ccacataacc tgagcatcat gtgcccagag 12240
cgttgggtgg tgtggtccca ttccttcctt ccagcggctt cttctggcca cctcaatgtc 12300
aggatgtcct gctcacatat caataccatt aaaacctgac ttctttccct gcactgttga 12360
agctccttct tgaggctcac attatggata taattttgat tctttcttca gtggtataga 12420
taactacttg taacctaaga acaacttggt gaaagtcctc taatacatta tttttaaaa 12480
aaacacaaat caatgagctc aacttattaa ctaactttca tctattcatt tttgagccat 12540
ccctgtctga ttgtgaatct ccatgattcc aacactctga gctggggata gtgcctacac 12600
aaaataaaaa gaagtggaaa attttcaaac atcagtttat gctgacaacc aggccataat 12660
aggtgctcaa ttactattga atgaatgaat gaaagttctg gccaggtacg gtggctcatg 12720
cctgtagtcc caacactttg ggaggccgag gcaggtggat cacttgaggt taggagttcg 12780
aaaccaacct gaccaacatg aagaaacctt atctctacca aaaaaatata aaaaaattac 12840
ccaggcatgg tggtgtatgc ctgtaatccc agctatttgg gaggctgagg caggaaaatc 12900
acttgaacct gagaggcgga ggttgcagtg agctgagatt gtgccactcc actccagcct 12960
gggcgacaga gtgagactcc gtcttactta aaaaaaaaaa aaagaaggtt ccaagaaaat 13020
tcatcttaag gtttatgtaa aaggaagatg atatttaaca tgattcatgg ccaagtacta 13080
atattacatt ataataatgt ttccaaataa cattatagat atgtttaaag acagtgtatt 13140
aggctgttct tgcattgctg taaagaaata cccaagactg ggtaatttat aaagaaaaga 13200
ggtttcattg gctcgtgttt ctgcaggctg tacaggaagc ttagtgctga catcacttgg 13260
ctgccggggg aacctcaggg agcttttact catggcagaa ggcaatgcgg gagcttgcat 13320
gtcacatggc aaaagcagga gcgagagaga gttggggggg aaggtgccac acactttta 13380
atgaccggct ctcacaataa ctcatgaaaa ctcactatca ggaagacagc actaaagcac 13440
aagggatccg accccatgat ccaaacacct cccaccaggc cccatctcca gcactgggga 13500
ttacaattca acatgagatc tgagtgtgga caaatatcca aactgtatca gtcaacagcg 13560
atcataatta gtcctgaata ggagtgcctt ttttttcttt tcttctccct tttcttttct 13620
acttcctcct ccttttccct ctcctcttca atctcctctt cattcctgta gcaccaaggg 13680
ttgaagcacc taacccgttt tggattgaga tgttctgatt gggcaatgaa cactgtccag 13740
aataaacaga aatccatttt gcactaagtg gctgcacaga ccctgcctca tgctaaatct 13800
agcacccaga tagtttaatg tttcaatgac tgaattacaa atatatcatc accttggatt 13860
tggcacttac aaatggctgt taatttggcc agaggtggtt gtttacaact tcaaatagga 13920
gactattcat aatttctgac gtgacatttt cctttcttta ttttactgta tgaaaatata 13980
atgaaatttc tcacaaaata tcactaaaaa gaaaagaaga agagtaggaa gcaaggttaa 14040
aatatttcta aaatataatt ttggtctttc tttttctccc ttccttcctc cgtccctctc 14100
tcctttcctc tctccctccc tccctccctc ccttcctcct ttccttgctt ccttccctcc 14160
ttctcttcct tctttttcaa gagatcaata acatttatta agaataagtt tcttaattat 14220
aacctttcag gtgataatag taacacagcc tgggcaacac aataagacct tgtttctaca 14280
aaaaatttaa aaattggcca gacatagtgg tgcatgacta attccagcta ctctggaggc 14340
tgaggcagga ggatggcttg agcccaggag ttggaggctg cagttagcca tgcttgtgcc 14400

actacactcc agcccgggca acwgggcaag actctgtatc taaaaacaac aacaacaaca 14460
ataatagaaa caggtttcct ttcccaagtt tggaaaatct ggtagtcttc ttaagcagcc 14520
atgagcataa agagaggatt gttcatacca caggtgttcc aggcataacg aaactgtctt 14580
tgtgtttagt tacaaggaga acatcatgcg cctctccagc cttcacaagg accgcccggt 14640
ggagccgctg gacctggccg tgttctgggt ggagtttgtg atgaggcaca agggcgcgcc 14700
acacctgcgc ccygcagccc acgacctcac ctggtaccag taccattcct tggacgtgat 14760
tggtttcctc ttggcygtcg tgctgacagt ggccttcatc acctttaaat gttgtgctta 14820
tggctaccgg aaatgcttgg ggaaaaaagg gcgagttaag aaagcccaca aatccaagac 14880
ccattgagaa gtgggtggga aataaggtaa aattttgaac cattccctag tcatttccaa 14940
acttgaaaac agaatcagtg ttaaattcat yttattctta ttaaggaaat actttgcata 15000
aattaatcag ccccagagtg ctttaaaaaa ttctcttaaa taaaaataat agactcgcta 15060
gtcagtaaag atatttgaat atgtatcgtg ccccctctgg tgtctttgat caggatgaca 15120
tgtgccattt ttcagaggac gtgcagacag gctggcattc tagattactt ttcttactct 15180
gaaacatggc ctgtttggga gtgcgggatt caaaggtggt cccacggctg cccctactgc 15240
aaatggcagt tttaatctta tcttttggct tctgcagatg gttgcaattg atccttaacc 15300
aataatggtc agtcctcatc tctgtcgtgc ttcataggtg ccaccttgtg tgtttaaaga 15360
agggaagctt tgtaccttta gagtgtaggt gaaatgaatg aatggcttgg agtgcactga 15420
gaacagcata tgatttcttg ctttggggaa aaagaatgat gctatgaaat tggtgggtgg 15480
tgtatttgag aagataatca ttgcttatgt caaatggagc tgaatttgat aaaaacccaa 15540
aatacagcta tgaagtgctg ggcaagttta ctttttttct gatgtttcct acaactaaaa 15600
ataaattaat aaatttatat aaattctatt taagtgtttt cactggtgtc gcatttattt 15660
cttgttaagt tgcattttct aattacaaaa gtaatgcatg attatgacag aaagtttgga 15720
aaatatagag gttcacacac acacgccttc attgcgtgtg catgcataaa tgcatgagaa 15780
aagaaaaata accagtaatc acatcgccca gaaataaccc cagttacaat tgtggcaaat 15840
acacatactt ataaatattg cagatatatt aagtatacct agtatttgct aacactcttt 15900
cttctactct gtcatgaaga ttctcccaag gtgtttttgt ataatattta attcatttc 15960
agtggccaag cagtattcta cttcatggat ataccaggat ttatttaacc ataacttctg 16020
gttggattca ctcttattat tttgtttaat taaaaaaaaa agacctcggc tgggcacagt 16080
ggctcatgcc tgtaatccca gcactttggg aggccgaggt gggtggatca cctaagatcg 16140
ggagtttgag accagcctgg ccaacatggc aaaaccccgt ctctactaaa aatacagaaa 16200
attagccggg tgtggttgcc agcacctgta attccagcta attgggaggc tgaggcagga 16260
gaattgcttg aaccggggtc aggggttcg gaggtcggag gttgcagtga gtccggatca 16320
tgccactgca ttccagcctg ggtgacacag ccagactctg tctcaaaaac aacaacaaca 16380
acaaaacaac aacaacaaca acaacaaaaa tctcactgga catcctagta gctaaggctt 16440
tccacatatt catgattact tctgttggaa agtgctttac aacaaattgc tagttgtctc 16500
agtctgggtt cccctgagat gaggattcaa gggccaggag tttatttagg aagtaaagga 16560
aacactgata gaggagtggc agagtgagaa ggggtgatgg tcatccacag ctggctctct 16620
tgtggtcaat cggagcttaa tcctgctggg tgactctggg agccagtgga gaaaagacac 16680
cccagactta tcccaatgag gaacacggct gttgggtgcc tgagtacttg cctcgtcagg 16740
gattgaaacg tactcccagg tagtagtaat ttctctgccc ttccattagg ccacaaaggg 16800
ggctctgaca gagagagctg acgagaaaaa acacacgccc ttgtcactga agaggtacac 16860
aggggatctg tgtggggcac cacctgcact gctaccctgg acaaatagct taagaaatcc 16920
ccacactgca tccccaaact tactatcagc gtgtgaggga gacaggttcc cacaccctca 16980
ttagcacaaa gtactatctt gaaaaagaaa gcctgtcagt ttgataggag aaaagcagga 17040
tcttgtttac aatgtgcttt tattattgtt attattagag attgtatttc ttttcaagct 17100
gatgagccgt ctgtgtttat tttttggagg ataccctttg cccactttcc tattggagtg 17160
tattaccctg aggatttggt aagagtgctt attgcattca ccagaatgtc ctttttgtca 17220
tttactgtat tttctctact ttttttttt tgccttgttt tactttttt gttttgtatt 17280

acaagcagaa gttttaaatt tgtaagcttc aaattggagc tggggtggtg cagagcgaag 17340
atttcagctg gttccctgac cccagctcca tctccttccc taggcagtgg ctggaacaca 17400
ttctgtccac tatttccctc tctacatcct tgaggctgtg cagtcacccc tcaactacgt 17460
tcaccctcct tcaaagccct tcctggtcca cccggggacc atctcccggc ctcactgccc 17520
ctagctcctt gacgccccaa cctctctcag ggaccccaag ttgccatgac ctccagccag 17580
ctcatgttca tttgcacctt cgtgtctgca gcactgaggc actcttgttt acaagtgaga 17640
gaacccaact cgggatacct taagcataaa cagtattttt gtaaggagac aggcttctga 17700
cgacgcgagg ctcatagcca ggcctgcgct gggtggagcc tccccttctc actcctgtcc 17760
ctgttgggtc agagtgccag ctttcctctt ccctctcctc cgggtctttt cggcccctca 17820
gtccccatat tcctctgccc tagctcccaa gatcccacaa gagacagact ggattctctc 17880
tggcctggag tgccaccttc ctgaaagtca gaatctgatt ggtccagctg gatcaggtgt 17940
cctctccctg tccaatcatc aatgccgaga gggattacgg agagaaaaac atggctccca 18000
ccatcccatt gctgtggctg cttgggccac gggagaggga accttgtgag ccgggcagaa 18060
tccaccctgt agagactgcc tctgggtgag tcatatggtt tggctgtgtc cccacccaaa 18120
tctcatcttg aattgtaatt cccattaccc ccatgtgtca ttagagggac ctggtgggaa 18180
gtgatttgaa ttatgggggc agttatctcc atgctatttg tgtgatagtg agttctcaca 18240
agatctgatg gttttatagg gggcttttcc ccctcttgct catacttcct cttgcctgcc 18300
accatgtaag atgtgccctt gctcctcctt caccttctgc catgattgtg aggcctcccc 18360
agccatgtgg aactgtgagt ccatgaaacc tctttttctt tataaattac ccagtcctgg 18420
gtatttcttc atagcagtat gaaaatggac taatacagtc agcttctgca caatatttcc 18480
attttcccac attatgtctt gggccttttg tgtatttaag ctcacaggat gctacgaata 18540
aagcgttttc ttatttcctg ggtagttccc atagaagtag tggtgcaacg tgccatagag 18600
tgacagcacc taagagaagc tgattttgtg agtggattgt gagttcaata ttgttgtcat 18660
aatcagaaaa aaatgtattt actttttttt tttttttttt tgagacggag tctcactctg 18720
ttgcccaggc tggagtgcag tggtgtgatc ttggctcact gcaacctccg cctcctgggt 18780
tcaagtgatt cttcctacct caccctcctg agtagctggg attacaggca catgccaccc 18840
caccacaccc ggctaatttt tgtatttttt agtagagaca gcgtttc 18887